Année : 2017 **Thèse N° : 71/ST2I**

École Nationale Supérieure d'Informatique et d'Analyse des Systèmes

Centre d'Études Doctorales en Sciences des Technologies de l'Information et de l'Ingénieur

THÈSE de Doctorat

Reconnaissance Statistique de la Parole Continue pour Voix Laryngée et Alaryngée

---

Présentée par:

Othman LACHHAB

Le samedi 15 avril 2017 à 10h à l'ENSET de Rabat.

Formation doctorale: Informatique

Structure de recherche: Équipe de recherche en Informatique et Télécommunications

Co-encadrant de thèse: **Dr. Joseph Di MARTINO, MC, LORIA, FRANCE.**

Jury:

Pr. Hassan QJIDAA, PES, FSDM, Fes.	Président et Rapporteur
Pr. Larbi BELLARBI, PES, ENSET, UM5, Rabat.	Rapporteur
Pr. Mohamed ET-TOLBA, PH, INPT, Rabat.	Rapporteur
Pr. Jamal El MHAMDI, PES, ENSET, UM5, Rabat.	Examinateur
Pr. Mounir AIT KERROUM, PH, ENCG, UIT, Kénitra.	Examinateur
Pr. El Hassane IBN ELHAJ, PES, INPT, Rabat.	Encadrant
Pr. Ahmed HAMMOUCH, PES, ENSET, Rabat.	Directeur de thèse

Dédicaces

A ceux que j'ai de plus chers

A ma très chère mère, symbole de douceur, de tendresse, d'amour et d'affection, grâce au sens du devoir et aux sacrifices immenses qu'elle a consentis, je suis parvenu à réaliser ce travail.

A mon très cher père, pour les sacrifices qu'il a consentis aussi pour mon éducation et pour l'avenir qu'il a su m'offrir.

A mes chers frères et sœurs

A ma chère sœur Fadoua, qui m'a toujours soutenu et encouragé durant tout mon parcours. Je suis chanceux de t'avoir à mes côtés.

A mon cher frère Hicham, qui m'a toujours encouragé et qui a toujours apprécié mon effort.

A mon cher frère et ami Yassir, qui m'a beaucoup aidé dans la vie et qui a toujours été présent à mes côtés.

A mon beau-frère Rachid, pour ses conseils et son encouragement durant ce travail.

A ma belle-sœur Imane, qui a toujours été une vraie sœur pour moi.

A la mémoire de mon très cher neveu Nizar, aucune dédicace, ni sentiment ne saurait exprimer l'amour, l'affection, l'estime et le dévouement que j'ai toujours eus pour toi. Jamais je ne t'oublierai, ton corps est parti mais ton âme est toujours présente avec nous.

A mes chers petits neveux et nièce Amjad, Yazid (Nizar 2), et Janna, aucune dédicace ne saurait exprimer tout l'amour que j'ai pour vous. Votre gaieté me comble de bonheur. Puisse Dieu vous garder, éclairer votre route et vous aider à réaliser à votre tour vos vœux les plus chers.

A tous mes enseignants à qui je dédie ce travail avec mes vifs remerciements et les expressions respectueuses de ma profonde gratitude.

Et enfin, **à tous mes amis**, Ali, Karim, Mehdi, Anis, Safouane, Abdellah, Amine, Tariq, Omar, Oussama, Hind, Ghita, Salma, Mouna et Zineb...

Je vous dédie en signe de reconnaissance ce travail qui n'a pu être accompli qu'avec vos encouragements et votre collaboration.

Othman

Remerciements

En premier lieu, je souhaite remercier chaleureusement mes deux encadrants, M. Elhassane Ibn Elhaj, Professeur de l'enseignement supérieur à l'Institut National des Postes et télécommunications (INPT), de Rabat, Maroc et M. Joseph Di Martino, Maître de Conférences au Loria (Université de Lorraine), Vandœuvre-lès-Nancy, France.

Effectivement je tiens à exprimer toute ma gratitude au Pr. Elhassane Ibn Elhaj pour ces années de soutien, pour ses précieux conseils scientifiques et pour son aide et sa capacité à simplifier les problèmes rencontrés dans le cadre du travail. Il m'a mis le pied à l'étrier après l'obtention de mon diplôme d'ingénieur et a toujours été présent lorsqu'il s'agissait de me donner un coup de main, ce qui fait de lui un encadrant/directeur idéal que tous les doctorants devraient avoir.

C'est également en toute sincérité que je remercie mon co-encadrant, Dr. Joseph Di Martino, de m'avoir si gentiment accueilli au sein de l'équipe Parole au Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA) et de de m'avoir consacré tout ce temps et toute cette énergie, toujours dans la bonne humeur. Sa réactivité et sa grande expérience dans le domaine de la reconnaissance automatique de la parole m'ont beaucoup apporté. Je lui suis donc très redevable de m'avoir permis de finaliser cette thèse dans de bonnes conditions.

Je tiens ensuite à exprimer ma gratitude à mon directeur de thèse M. Ahmed Hammouch, Professeur de l'enseignement supérieur à l'École Normale Supérieur de l'Enseignement Technique (ENSET) et directeur du Centre National pour la Recherche Scientifique et Technique (CNRST). Je le remercie pour avoir accepté de diriger mes travaux de recherche et aussi pour le suivi et l'implication inconditionnelle portés à cette thèse malgré son emploi du temps chargé.

Je remercie également M. Hassan Qjidaa d'avoir accepté de présider et rapporter mon travail de thèse et pour ces remarques judicieuses concernant ce manuscrit.

Je ne manquerai pas de remercier M. Larbi Bellarbi et M. Mohamed Et-Tolba, d'avoir accepté de juger la qualié de mon travail en tant que rapporteurs.

Je tiens aussi à remercier MM. Jamal El MHamdi et Mounir Ait Kerroum pour avoir examiné mon manuscrit avec précision et pour avoir soulevé les bonnes questions.

Au cour de cette thèse, j'ai bénéficié d'une bourse d'excellence octroyée par le CNRST dans le cadre du programme des bourses de recherche initié par le ministère de l'éducation national de l'enseignement supérieur, de recherche scientifique et de la formation des cadres. Durant mes séjours en France, j'ai bénéficié d'une bourse de mobilité dans le cadre du projet de recherche Inria Euro-Méditerranéens 3+3 Oesovox et du programme Européen Coadvise FP7. Je tiens ainsi à exprimer toute ma gratitude aux comités de sélection d'Inria et du FP7.

Durant toutes ces années, j'ai eu l'occasion de rencontrer de nombreuses personnes, dans un cadre purement professionnel ou simplement amical. A leur façon, ils ont tous contribué à mon apprentissage. Je suis reconnaissant envers chacune de ces personnes.

Je tiens à remercier tous mes collègues du laboratoire informatique de l'INPT ainsi que les membres de l'équipe Parole du laboratoire LORIA.

Finalement, je souhaite remercier vivement tous les étudiants avec qui j'ai eu la chance de travailler.

Table des matières

Liste des abréviations et notations

ACP Analyse en Composantes Principales

API Alphabet Phonétique International

AR Auto Régressif

CELP Code-Excitated Linear Prediction

CF Cepstre de Fourier

DFW Dynamic Frequency Warping

DTW Dynamic Time Warping

F0 La fréquence fondamentale

FD-PSOLA Frequency Domain PSOLA

FPSD French Pathological Speech Database

GMM Gaussian Mixture Model

HLDA Heteroscedastic Linear Discriminant Analysis

HMM Hidden Markov Model

HTK Hidden Markov Model Toolkit

ISE2D Iterative Statistical Estimation Directly from Data

LBG Algorithme de Linde Buzo et Gray

LDA Linear Discriminant Analysis

LPC Linear Predictive Coding

LPCC Linear Prediction Cepstral Coefficients

LSF Linear Spectral Frequency

MAP Maximum A Posteriori

MELP Mixed-Excitation Linear Prediction

MFCC Mel Frequency Cepstral Coefficients

MLE Maximum Likelihood Estimation

MMC Modèles de Markov Cachés

MMI Maximum Mutual Information

PER Phone Error Rate

PLP Perceptual Linear Prediction

PPL PerPLexité

PSOLA Pitch-Synchronous OverLap-Add

QV Quantification Vectorielle

RAP Reconnaissance Automatique de la parole

RLM Régression Linéaire Multivariée

SAMPA Speech assessment Methods Phonetic Alphabet

SoX Sound eXchange

SRAP Système de Reconnaissance Automatique de la Parole

STRAIGHT Speech Transformation and Representation using Adaptative Interpolation of weiGHTed spectrum

TCD Transformée en Cosinus Discrète

TD-PSOLA Time-Domain PSOLA

TFD Transformation de Fourier Discrète

TTS Text-To-Speech

Liste des figures

Liste des tableaux

Résumé

La Reconnaissance Automatique de la Parole (RAP) demeure depuis toujours un défi scientifique. Au cours de ces dernières années de grands efforts de recherche ont été concrétisés, afin de développer des systèmes d'aide et des solutions permettant d'effectuer certaine tâches jusqu'ici réservées aux humains. La parole est un mode de communication naturel, et un moyen facile pour échanger des informations entre humains. Une personne laryngectomisée, n'a pas la capacité de parler normalement puisqu'elle est dépourvue de ses cordes vocales suite à une ablation chirurgicale du larynx. Ainsi, le patient perd toute possibilité de communication avec une voix laryngée. Néanmoins, la rééducation avec un orthophoniste lui permet d'acquérir une voix de substitution dite "œsophagienne". Contrairement à la parole laryngée (normale), cette parole œsophagienne (alaryngée) est rauque, faible en énergie et en intelligibilité ce qui la rend difficile à comprendre.

L'objectif de cette thèse est la réalisation d'un système de reconnaissance automatique de la parole œsophagienne (alaryngée). Ce système devrait être en mesure de restituer, la plus grande partie des informations phonétiques contenues dans le signal de la parole œsophagienne. Cette information textuelle fournie par la partie décodage de ce système pourra être utilisée par un synthétiseur texte-parole (Text-To-Speech) dans le but de reconstruire une voix laryngée. Un tel système permettrait aux personnes laryngectomisées, une communication orale plus facile avec d'autres personnes.

Notre première contribution est relative au développement d'un système de reconnaissance automatique de la parole laryngée en utilisant des modèles de Markov cachés. Les rares corpus de parole œsophagienne existants, ne sont pas dédiés à la reconnaissance, à cause d'un manque de données (souvent quelques dizaines de phrases sont enregistrées). Pour cette raison, nous avons conçu notre propre base de données dédiée à

la reconnaissance de la parole œsophagienne contenant 480 phases prononcées par un locuteur laryngectomisé. Dans une seconde partie, le système de reconnaissance de la parole laryngée créé a été adapté et appliqué à cette parole œsophagienne. Notre dernière contribution au sujet de cette thèse concerne la réalisation d'un système hybride (correction = conversion + reconnaissance) fondé sur la conversion de la voix en projetant les vecteurs acoustiques de la parole œsophagienne dans un espace moins perturbé et relatif à la parole laryngée. Nous montrons que ce système hybride est capable d'améliorer la reconnaissance de cette parole alaryngée.

Mots clés

Système de Reconnaissance Automatique de la Parole (SRAP), Conversion de Voix (CV), Modèles de Markov Cachés (MMC), Modèles de Mélange de Gaussiennes (MMG), Reconnaissance automatique de la parole œsophagienne, Correction et amélioration de la parole œsophagienne.

Abstract

Automatic Speech Recognition (ASR) has always been a scientist challenge. Many research efforts have been made over recent years to offer solutions and aiding systems in order to carry out various tasks previously dedicated only to humans. Speech is considered the most natural mode of communication, and an easy way for exchanging information between humans. A laryngectomee person lacks the ability of speaking normally because he/her lost his/her vocal cords after a surgical ablation of the larynx. Thus, the patient loses the phonation ability. Only a reeducation by a speech therapist allows this person to provide a new substitution voice called “esophageal”. Unlike laryngeal speech (normal), esophageal speech (alaryngeal) is hoarse, weak in intensity and in intelligibility which makes it difficult to understand.

The goal of this thesis is the implementation of an automatic esophageal speech (alaryngeal) recognition system. This system should be able to provide most of the phonetic information contained in the esophageal speech signal. The decoding part of this system connected to a text-to-speech synthesizer should allow the reconstruction of a laryngeal voice. Such a system should permit laryngectomees an easier oral communication with other people.

Our first contribution concerns the development of an automatic laryngeal speech recognition system using hidden Markov models. The few existing corpora of esophageal speech, are not dedicated to recognition, because of a lack of data (only a few dozen sentences are registered in practice). For this reason, we designed our own database dedicated to esophageal speech recognition containing 480 sentences spoken by a laryngectomee speaker. In the second part, our devoted laryngeal speech recognition system has been adapted and applied to this esophageal speech. Our last contribution of this thesis concerns the realization of a hybrid system (correction = conversion + recognition) based

on voice conversion by projecting the acoustic feature vectors of esophageal speech in a less disturbed space related to laryngeal speech. We demonstrate that this hybrid system is able to improve the recognition of alaryngeal speech.

Keywords

Automatic Speech Recognition System (ASRS), Voice Conversion (VC), Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), Automatic esophageal speech recognition, Correction and enhancement of esophageal speech.

Introduction Générale

Problématique

La parole est certainement le mode de communication le plus naturel que les humains utilisent pour interagir les uns avec les autres. Ceci, peut être justifié par le fait que le signal vocal de la parole permet la transmission intelligible d'une importante quantité d'informations. Une personne laryngectomisée, n'a pas la capacité de parler normalement puisqu'elle est dépourvu de ses cordes vocales suite à une ablation chirurgicale du larynx. Ainsi le patient perd toute possibilité de communication avec une voix laryngée. Après la chirurgie, la phonation est impossible et certains patients peuvent renoncer à toute tentative de communication orale en raison du bouleversement physique et mental causé par l'acte chirurgical. Dès la période post-opératoire, le patient doit trouver de nouveaux moyens de communication afin de pallier la perte de ses cordes vocales et donc l'absence de voix laryngée. Seule la rééducation avec un orthophoniste lui permet d'acquérir une voix de substitution dite "œsophagienne". Contrairement à la parole laryngée, cette parole alaryngée (œsophagienne) est caractérisée par un bruit élevé, une faible intelligibilité et une fréquence fondamentale instable. Toutes ces caractéristiques permettent de produire une voix rauque, grinçante et non naturelle, difficile à comprendre. Pour ces raisons plusieurs approches ont été proposées pour améliorer la qualité et l'intelligibilité de cette parole œsophagienne. Citons par exemple : le remplacement du voisement humain par des signaux d'excitation artificiels [Loscos et Bonada, 2006], l'amélioration des caractéristiques spectrales à l'aide d'une synthèse de voix par formants [Matui et collab., 1999], la réduction du bruit de fond basé sur un masquage auditif [Liu et collab., 2006].

D'autres progrès ont été réalisés visant l'amélioration de la voix œsophagienne grâce aux techniques dites de "conversion de la voix". Généralement, la conversion de la voix est proposée dans le but de transformer la voix laryngée d'un locuteur source en celle d'un

locuteur cible. Dans [NING et YINGYONG, 1997], [DOI et collab., 2014] et [TANAKA et collab., 2014], des systèmes de correction de la voix alaryngée ont été développés, fondés sur la conversion vocale en transformant la voix du locuteur source (alaryngée) en une voix cible (laryngée). Tous ces systèmes correctifs utilisent un module de re-synthèse vocale pour reconstruire la parole convertie. Cependant, il est difficile de compenser les différences existantes au niveau des paramètres acoustiques de la parole alaryngée (par rapport à ceux de la parole laryngée) en utilisant une re-synthèse vocale après la conversion. Ceci, peut être expliqué par le fait que les signaux excitatifs calculés sont peu réalistes.

De nos jours, l'évaluation de la parole alaryngée est sortie du simple cadre de la recherche clinique et intéresse les laboratoires de recherche en traitement du signal et de la parole. L'évaluation par des jugements de perception est une méthode très coûteuse en temps et en ressources humaines et ne peut être planifiée régulièrement. C'est pour cette raison que l'évaluation et le décodage de la parole alaryngée par une méthode instrumentale devient une priorité. L'objectif de la reconnaissance automatique de la parole est d'extraire l'information lexicale contenue dans un signal de parole par le biais d'un système informatique. Cette technologie peut être utilisée avec succès sur la parole œsophagienne pour décoder l'information phonétique afin de comprendre le discours et faciliter la communication d'une personne laryngectomisée. C'est donc ce défi que nous nous envisageons de relever au cours de cette thèse.

Contributions

Notre première contribution dans cette thèse réside dans la création de notre propre système de reconnaissance automatique de la parole laryngée nommé SPIRIT [LACHHAB et collab., 2012]. Ce système est basé sur les travaux de recherche effectués au sein de l'équipe Parole de Nancy sur la reconnaissance de phonèmes isolés en utilisant la base de données TIMIT [GAROFOLO et collab., 1993]. Nous avons réussi à adapter et appliquer ces méthodes à la reconnaissance de phonèmes connectés indépendante du locuteur. Une modélisation de la durée d'émission des modèles phonétiques HMM (Hidden Markov Model) basée sur une distribution gaussienne a été proposée pour améliorer le taux de décodage de la parole de ce système. Nous avons implémenté aussi deux autres systèmes de reconnaissance automatique de la parole à l'aide de la plate-forme HTK (Hidden

Markov Model Toolkit [YOUNG et collab., 2006] : l'un basé sur des modèles phonétiques indépendants du contexte (monophones) et l'autre plus performant fondé sur une modélisation triphone des modèles phonétiques qui tient compte du contexte phonétique gauche et droit. En plus, la transformation discriminante HLDA (Heteroscedastic Linear Discriminant Analysis) [KUMAR et ANDREOU, 1998] a été appliquée sur les vecteurs acoustiques pour améliorer l'information discriminante entre les classes phonétiques et a permis ainsi, une augumentation significative du taux de reconnaissance phonétique.

Notre deuxième contribution est relative à la construction de notre propre base de données de la parole œsophagienne. Ce corpus intitulé FPSD "French Pathological Speech Database" [LACHHAB et collab., 2014] est dédiée à la reconnaissance automatique de la parole œsophagienne. Celui-ci contient 480 phrases prononcées par un locuteur laryngectomisé qui a acquis la voix œsophagienne après une rééducation vocale. Ces 480 phrases ont été segmentées manuellement en mots et en phonèmes afin de faciliter l'apprentissage et le décodage du système de Reconnaissance Automatique de la Parole (RAP). Le système de reconnaissance monpohone de la parole laryngée a été ensuite adapté à la parole œsophagienne permettant ainsi d'élaborer une technique objective [LACHHAB et collab., 2014] pour l'évaluation et le décodage de cette parole.

Notre troisième contribution réside dans la réalisation d'un système hybride [LACHHAB et collab., 2015] pour la correction des distorsions présentes dans les vecteurs acoustiques de la parole œsophagienne. Ce système hybride de correction est basé sur la conversion de la voix en projetant les vecteurs acoustiques de la parole œsophagienne dans un espace plus "propre" relatif à la parole laryngée. Nous n'utilisons pas un algorithme de re-synthèse vocale pour reconstruire les signaux de la parole convertie, parce que les vecteurs acoustiques convertis sont utilisés directement comme entrées par le système de reconnaissance monophone. Ce système hybride intègre aussi la transformation HLDA des vecteurs acoustiques et permet d'améliorer le décodage de la parole œsophagienne.

Organisation de la thèse

Cette thèse est organisée en quatre chapitres. Nous présentons dans le premier chapitre des généralités sur le signal de la parole ainsi que l'architecture fonctionnelle d'un système RAP. Nous découvrirons les modèles phonétiques et de langages couramment utilisés ainsi qu'une description précise des algorithmes d'apprentissage Baum-Welch et de décodage Viterbi.

Nous détaillerons dans le deuxième chapitre la mise en œuvre de nos trois systèmes de reconnaissance automatique de la parole laryngée ainsi que la transformation discriminante HLDA des vecteurs acoustiques. Nous évaluerons ces systèmes à l'aide de la base de données TIMIT.

Le troisième chapitre présentera les différents types de voix alaryngées et la cause des distorsions de ce type de signaux vocaux. Ensuite, nous exposons les caractéristiques de la parole de substitution œsophagienne. Nous dérirons les étapes de la conception de notre corpus FPSD dédiée à la reconnaissance de la parole œsophagienne. Nous conclurons ce chapitre par l'adaptation du système de reconnaissance monophone de la parole laryngée à la parole œsophagienne.

Nous nous focaliserons dans le quatrième et dernier chapitre sur les techniques correctives de la parole œsophagienne. Nous détaillerons aussi la mise en œuvre de notre système hybride de correction capable d'améliorer la reconnaissance automatique de la parole œsophagienne.

Contexte : laboratoires de recherche

Ce travail de doctorat, a été financé par le Centre National pour la Recherche Scientifique et Technique (CNRST) et par le projet Européen IRSES-COADVISE (FP7) et s'inscrit dans le cadre des projets de recherche Inria Euro-Méditerranéens 3+3 M06/07 Larynx et M09/02 Oesovox. Il a été réalisé au sein de trois laboratoires :

- Laboratoire de Recherche en Génie Electrique (LRGE), au sein de l'équipe de recherche en Informatique et Télécommunications de L'Ecole Normale Supérieure de l'Enseignement Technique (ENSET), Rabat, Maroc.
- Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA), au sein de l'équipe Parole, Centre de Recherche Inria Nancy - Grand Est, Villers-lès-Nancy, France.
- Laboratoire Informatique de l'Institut National de Postes et Télécommunications (INPT), Rabat, Maroc.

Chapitre 1

État de l'art

« *Tout ce que je sais, c'est que je ne sais rien.* »

Socrate

1.1 Introduction

L'objectif de la Reconnaissance Automatique de la Parole (RAP), est d'extraire l'information textuelle contenue dans un signal de la parole à l'aide d'un logiciel informatique. Différentes approches ont été développées pour réaliser cette tâche complexe. Actuellement, la technique la plus performante est fondée sur une modélisation statistique des sons élémentaires en utilisant les modèles de Markov cachés (Hidden Markov Models - HMMs) : l'étude et le développement de cette technique est le sujet principal de cette thèse ; mais nous tenons à préciser pour être complètement objectif qu'une autre approche, fondée sur une modélisation neuronale, est très étudiée à l'heure actuelle et a de fortes chances de supplanter les méthodes probabilistes avec en contrepartie un accroissement considérable du temps de calcul surtout pour la phase d'apprentissage. Cette dernière approche ne sera pas explicitée dans ce manuscrit.

La grande redondance du signal de la parole ne lui permet pas d'être exploité directement dans son état initial. En effet, l'extraction des paramètres qui sont dépendants de l'information linguistique est nécessaire.

Généralement, les vecteurs cepstraux MFCC (Mel Frequency Cepstral Coefficients) [DAVIS et MERMELSTEIN, 1980] sont les paramètres les plus couramment utilisés dans le domaine de la reconnaissance de la parole. Ceux-ci tiennent compte de connaissances acquises sur la production, la perception et la variabilité du signal de la parole.

Nous allons donc présenter dans ce chapitre, le problème lié à la reconnaissance de la parole, les différents concepts pour construire un système RAP que sont la paramétrisation, la modélisation acoustique et la modélisation linguistique.

Les algorithmes d'apprentissage et de reconnaissance (décodage) mettant en œuvre les modèles HMMs seront aussi détaillés dans ce chapitre. Nos travaux de recherche sont directement liés à ces concepts.

1.2 Complexité du signal de la parole

Le défi sous-jacent à la technologie de reconnaissance vocale est la grande complexité particulière existante dans le signal de la parole. En effet, plusieurs facteurs sont à l'origine

de cette complexité, en particulier la redondance, la continuité et les effets de coarticulation, et l'ample variabilité intra et inter-locuteurs. Toutes ces caractéristiques doivent être prises en compte lors de la création d'un système RAP.

1.2.1 Redondance

Le signal de parole est redondant car il transporte énormément d'informations (des informations liées au locuteur, son état émotionnel, sa prosodie, son timbre, l'information lexicale, etc...) ; toutes ces informations ne sont pas forcément utiles pour faire de la reconnaissance automatique de la parole. Ainsi, il est important d'extraire les caractéristiques qui dépendent uniquement du message linguistique. L'analyse ou paramétrisation a pour objectif d'extraire seulement les paramètres pertinents pour la tâche envisagée (RAP) et ainsi réduire la redondance du signal de la parole.

1.2.2 Continuité et coarticulation

Lorsque l'on entend parler une langue connue, on perçoit une continuité de mots, qui peuvent à leur tour être décrits comme une suite de sons élémentaires appelés phonèmes. Le phonème est une unité sonore distinctive minimale de la chaîne parlée, qui permet de différencier 2 mots (lampe et rampe /l/ et /r/ sont 2 phonèmes distincts en français). La langue française peut être représentée au minimum par une trentaine de phonèmes. Malheureusement, l'analyse du signal vocal ne permet pas de déceler les marques de séparation entre mots successifs et aussi entre les phonèmes successifs à l'intérieur des mots. La production de la parole se fait par un flux continu de phonèmes profondément influencés par les sons qui les succèdent ou qui les précèdent, créant ainsi des phénomènes de coarticulation.

1.2.3 Variabilité

Un mot n'est jamais prononcé deux fois exactement de la même façon, même par le même locuteur (variabilité intra-locuteur) ou par des locuteurs différents (variabilité inter-locuteur). La différence au niveau du signal vocal entre deux prononciations d'un

même énoncé à contenu phonétique égal peut être causée par plusieurs facteurs :

- **Variabilité intra-locuteur :**
  - → L'état physique (rhume ou fatigue).
  - → Les émotions du locuteur.
  - → Le rythme d'élocution et l'intensité du discours (voix normale, voix criée, voix chuchotée).
- **Variabilité inter-locuteur :**
  - → Le timbre.
  - → Le sexe et l'âge du locuteur : homme, femme, enfant, adulte, vieillard.
  - → La prononciation régionale dans un milieu social (les accents).

La reconnaissance de la parole continue est donc très imparfaite, particulièrement en fonctionnement multilocuteurs.

1.3 Architecture d'un système RAP

Le but d'un système de reconnaissance automatique de la parole est de fournir la transcription textuelle d'un signal audio fourni en entrée. Il peut être décomposé en cinq modules, comme illustré dans la figure 1.1 :

A) **Un module de paramétrisation et de traitement du signal :** permet d'extraire l'information utile à la caractérisation de son contenu linguistique en réduisant la redondance du signal de la parole. Le signal sonore brut est converti en une séquence de vecteurs acoustiques adaptée à la reconnaissance.

B) **Des modèles acoustiques :** modélisant un ensemble réduit d'unités de sons élémentaires d'une langue donnée. C'est unités acoustiques sont plus petits que les mots par rapport au nombre d'échantillons. Ce sont des modèles phonétiques statistiques (HMMs) estimés à l'aide d'une grande quantité de données de parole.

C) **Un modèle lexical :** fourni la transcription de mots de la langue modélisée par un simple dictionnaire phonétique. Les plus développés sont construits à partir des automates probabilistes, capables de représenter chaque mot d'un dictionnaire par une probabilité.

D) **Un module de langage :** introduit la notion de contraintes linguistiques par un modèle statistique utilisant une grande base de données textuelles pour estimer les probabilités d'une suite de phonèmes, de manière automatique. Il permet de guider le décodeur vers les suites de mots les plus probables.

E) **Un module de décodage :** consiste à sélectionner, parmi l'ensemble des phrases possibles, celle qui correspond le mieux à la phrase prononcée. Le décodage de la parole s'effectue à l'aide de tous les modules déjà présentés.

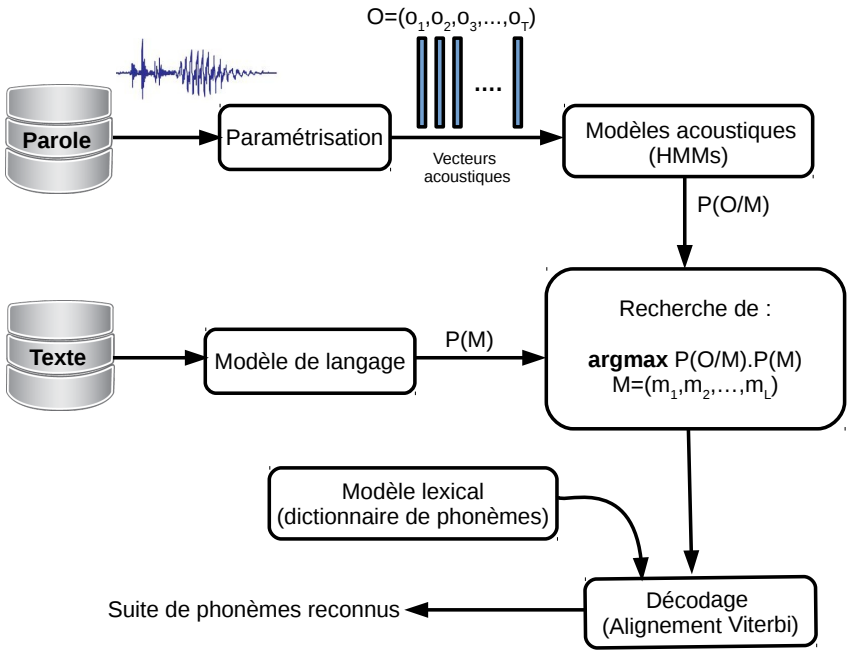

FIGURE 1.1: *Architecture d'un système de reconnaissance automatique de la parole*

Les systèmes de RAP continue qui ont nécessité le plus d'effort de recherche jusqu'à présent sont fondés sur une méthode statistique [JELINEK, 1976] basée sur les modèles de Markov cachés. Après l'étape de paramétrisation, nous obtenons une séquence O de T observations (vecteurs), $O = (o_1, o_2, \ldots, o_T)$. Effectuer la reconnaissance d'une phrase revient à déterminer la séquence de phonèmes $\tilde{M} = m_1 \ldots m_n$ qui maximise la probabilité que cette séquence corresponde à la suite d'observations O. Ce problème peut s'écrire ainsi :

$$\tilde{M} = \arg\max_{M} P(M/O) \tag{1.1}$$

Toutefois, il est impossible de calculer directement la probabilité P(M/O). Cependant, en utilisant la règle de Bayes (équation 1.2), il est possible d'écrire la probabilité qu'une séquence de phonèmes correspond aux observations données comme :

$$P(M/O) = \frac{P(O/M).P(M)}{P(O)} \tag{1.2}$$

Par cette nouvelle formulation, nous obtenons l'expression du problème en fonction de trois autres probabilités :

- P(O/M) : La probabilité d'observer la séquence O des vecteurs acoustiques sachant la suite de phonèmes M. Cette probabilité est estimée par les modèles acoustiques (module B).
- P(M) : La probabilité a priori d'observer la suite de phonèmes M, indépendamment du signal. Elle est déterminée par le modèle de langage (module D).
- P(O) : La probabilité d'observer la séquence de vecteurs acoustique O. Elle est identique pour chaque suite de phonèmes (P(O) ne dépend pas de M). Elle n'est pas utile et peut donc être ignorée.

Alors l'équation 1.1 est simplifiée par l'équation 1.3 qui ne dépend plus que des probabilités acoustiques et linguistiques :

$$\tilde{M} = \arg\max_{M} P(O/M).P(M) \tag{1.3}$$

Cette méthode statistique permet de représenter, de manière élégante, les niveaux acoustiques et linguistiques dans le même processus de reconnaissance. Nous décrivons dans les sections suivantes chaque module du système de RAP continue.

1.4 Paramétrisation et traitement du signal

La grande redondance et variabilité du signal de la parole ne lui permet pas être exploité directement dans son état initial par un système RAP. Il est donc essentiel de convertir ce signal en paramètres acoustiques qui sont dépendants de l'information linguistique.

Divers méthodes de paramétrisation ont été proposées, les plus utilisées en fonction du domaine d'analyse sont :

- Les **MFCC** (Mel Frequency Cepstral Coefficients) [DAVIS et MERMELSTEIN, 1980].
  - → Domaine cepstral

⊛ Les **PLP** (Perceptual Linear Prediction) [HERMANSKY, 1990].

→ Domaine spectral

⊛ Les **LPCC** (Linear Prediction Cepstral Coefficients) [MARKEL et GRAY, 1976].

→ Domaine temporel

Pour notre étude on s'intéressera surtout à la représentation MFCC qui est décrite ci-dessous.

1.4.1 Coefficients Mel-Cepstraux

Les principales étapes de calcul des coefficients cepstraux MFCC sont décrites dans la figure 1.2. Le signal de la parole est variant au cours du temps. Pour cette raison, il doit être divisé en trames de faible durée (typiquement 20 à 30 ms) où le signal sonore peut être considéré comme quasi-stationnaire, avec un pas de décalage entre deux trames successives de l'ordre de 10 ms. Un vecteur cepstral est extrait pour chaque trame. Le signal de la parole S_n est pré-accentué à l'instant n pour relever les hautes fréquences par l'équation 1.4, pour une valeur classique α de 0.97 (α peut prendre une valeur comprise entre 0.9 et 1).

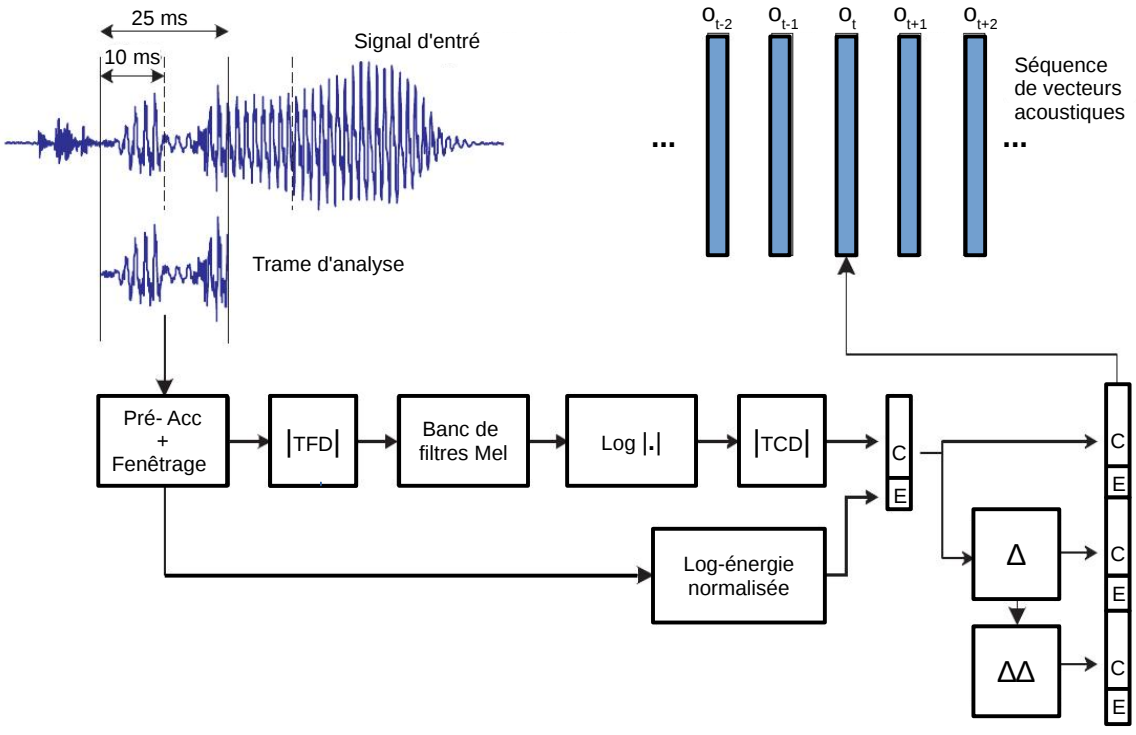

FIGURE 1.2: *Module de paramétrisation par la représentation MFCC*

$$S_n = S_n - \alpha S_{n-1} \tag{1.4}$$

Ensuite, on applique sur chaque trame une fenêtre de Hamming [HARRIS, 1978] pour rendre proche de zéro les extrémités de la trame temporelle.

$$S_n = S_n.[0.54 - 0.46\cos(2\pi\frac{n}{N-1})], \quad 0 \leq n \leq N-1 \tag{1.5}$$

Les n premiers coefficients cepstraux C_k (en général n est choisi entre 10 et 15) sont calculés directement à partir du logarithme des énergies m_i sortant d'un banc de F filtres en échelle de fréquences non linéaire Mel ou Bark. Cette opération est appelée transformation en cosinus discrète (DCT).

$$C_k = \sum_{i=1}^{F} \log m_i \cos[\frac{\pi k}{F}(i-0.5)], \quad 1 \leq k \leq n \tag{1.6}$$

Le coefficient C_0 représentant l'énergie moyenne de la trame du signal est souvent éliminé. Il est éventuellement remplacé par le logarithme de l'énergie total E calculée par l'équation 1.7 suivante :

$$E = \log \sum_{n=0}^{N-1} S_n^2 \tag{1.7}$$

Qui est normalisé comme ceci :

$$\bar{E} = 0.1(E - E_{max}) + 1.0 \tag{1.8}$$

Où E_{max} représente le maximum de E calculé sur tout le signal analysé.

1.4.2 Coefficients différentiels

Les coefficients MFCC sont généralement considérés comme des coefficients statiques. Ces paramètres initiaux, seront ensuite traités comme une séquence d'observations par un HMM en tant que modèle acoustique (voir la section suivante 1.5). Ces observations sont conditionnellement indépendantes et l'information dynamique locale dans chaque état, est perdue. Pour garder cette information, on étend ces paramètres initiaux avec leurs dérivées (temporelles) [FURUI, 1986] premières et secondes.

Soit $C(t)$ le vecteur cepstral de la trame t, alors le vecteur différentiel d'ordre 1 correspondant $\Delta C(t)$ (vitesse) est calculé à l'aide d'une fenêtre d'analyse de cinq trames ($N_\tau = 2$) en utilisant l'équation suivante :

$$\Delta C(t) = \frac{\sum_{i=1}^{N_\tau} i(C_{t+i} - C_{t-i})}{2\sum_{i=1}^{N_\tau} i^2} \tag{1.9}$$

La même formule 1.9 est appliquée sur les coefficients delta pour obtenir l'accélération ($\Delta\Delta$ ou dérivée seconde). Les dérivées de l'énergie sont calculées aussi de la même façon.

L'application de ces coefficients différentiels améliore sensiblement les performances des systèmes RAP basées sur les modèles HMM [LEE et HON, 1989][WILPON et collab., 1993][LAMEL et GAUVAIN, 1993]. Une amélioration de 6% du taux de reconnaissance phonétique est obtenue par le système SPHINX [LEE et collab., 1990] sur la base de données TIMIT[GAROFOLO et collab., 1993].

1.5 Modélisation acoustique

La modélisation du signal de la parole est effectuée sur un ensemble réduit d'unités sonores, plus courtes que les mots, typiquement les phonèmes. Les unités acoustiques les plus utilisées en reconnaissance de la parole continue sont les phonèmes dépendants du contexte. Lorsque le phonème est dépendant du contexte gauche et droit (phonème précèdent et phonème suivant), on parle de triphone. Dans la littérature, plusieurs modélisations ont été proposées pour représenter les unités acoustiques. Parmi les plus fréquentes, on trouve les réseaux de neurones [ROBINSON et FALLSIDE, 1991][ROBINSON, 1994][TEBELSKIS, 1995], les réseaux bayesiens [MING et SMITH, 1998; ZWEIG et RUSSELL, 1999], les machines à support vectoriel [VAPNIK, 1998]. La solution la plus utilisée depuis déjà une trentaine d'années est fondée sur les modèles de Markov cachés (Hidden Markov Model - HMM) [BAKER, 1975][JELINEK, 1976][RABINER, 1989]. Nous détaillons cette technique dans la section suivante.

1.5.1 Modèle de Markov caché

Un modèle de Markov caché correspond à un automate probabiliste à N états comportant deux processus. Un processus caché de transition d'état, car l'état dans lequel se trouve celui-ci à l'instant t n'est pas connu (caché). Le deuxième est un processus d'émission des observations (vecteurs acoustiques). Dans le cas d'un processus markovien (d'ordre 1), la probabilité de passer de l'état i à l'état j à l'instant t en émettant l'observation o_t ne dépend pas des états parcourus aux instants précédents.

Dans le cas de la parole continue, chaque phonème doit être modélisé par un modèle de Markov caché, gauche-droite à cinq états mais trois seulement d'entre eux sont émetteurs. L'état initial et l'état final ont pour objectif de servir uniquement à la connexion des modèles en parole continue sans émettre d'observation. Les transitions entre les états sont irréversibles, de la gauche vers la droite. La figure 1.3, illustre la topologie et le type d'HMM utilisé.

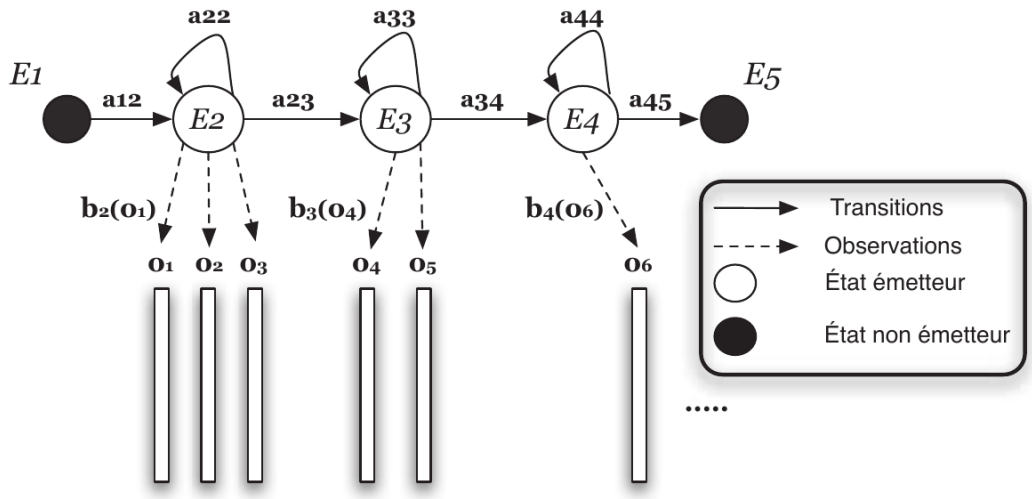

FIGURE 1.3: *La topologie d'un modèle phonétique HMM indépendant du contexte*

Un HMM est caractérisé par l'ensemble des paramètres :

- La séquence d'états $\mathscr{E} = (E_1, E_2, \ldots, E_N)$, (N = 5 dans notre cas)
- La séquence d'observations $O = (o_1, o_2, \ldots, o_T)$ associée à la séquence d'états $\mathscr{E}$.
- $\boldsymbol{\pi_i}$ la probabilité initiale, c'est à dire la probabilité d'être dans l'état i à l'instant initial.

- ⊛ $\mathbf{A} = (\mathbf{a_{ij}})$: la matrice de transition entre états, a_{ij} représente la probabilité de transition pour aller de l'état i à l'état j.

$$a_{ij} = \mathrm{P}(\mathrm{E}_t = j / \mathrm{E}_{t-1} = i), \quad \forall i, j \in \{1, \mathrm{N}\}$$

  Cette matrice n'est pas pleine et on peut soit se déplacer à droite soit rester dans l'état courant. Les coefficients de cette matrice A doivent vérifier la propriété suivante : $\forall i, \sum_{j=1}^{\mathrm{N}} a_{ij} = 1$

- ⊛ $\mathbf{B} = \mathbf{b_i}(\mathbf{o_t})$: La probabilité d'observer le vecteur o_t sachant que le processus Markovien est dans l'état i.

$$b_i(o_t) = \mathrm{P}(o_t / \mathrm{E}_t = i), \quad \begin{cases} \forall i \in \{1, \mathrm{N}\} \\ \forall t \in \{1, \mathrm{T}\} \end{cases}$$

La probabilité d'émission $b_i(o_t)$ des observations continues O_t est généralement calculée par une somme pondérée de G_i gaussiennes $\mathcal{N}(\mu, \Sigma)$ appelé aussi modèle de mélange de gaussiennes (Gaussian Mixture Model- GMM), chaque gaussienne est caractérisée par un vecteur moyen μ_{ik} et une matrice de covariance Σ_{ik}.

La probabilité d'émission $b_i(o_t)$ est alors définie par la formule suivante :

$$\begin{aligned} b_i(o_t) &= \sum_{k=1}^{\mathrm{G}_i} w_{ik} \mathcal{N}(o_t, \mu_{ik}, \Sigma_{ik}), \quad \sum_{k=1}^{\mathrm{G}_i} w_{ik} = 1 \\ &= \sum_{k=1}^{\mathrm{G}_i} \frac{w_{ik}}{\sqrt{(2\pi)^d |\Sigma_{ik}|}} \exp(-0.5(o_t - \mu_{ik})' \Sigma_{ik}^{-1} (o_t - \mu_{ik})) \end{aligned} \tag{1.10}$$

Où G_i représente le nombre de gaussiennes de l'état i, w_{ik} représente le poids de pondération de la $k^{ième}$ gaussienne dans l'état i, pour laquelle o_t représente le vecteur d'observation à d coefficients.

D'autre types de densités de probabilités sont possibles, comme par exemple une représentation paramétrique : le Laplacien ou l'erreur de prédiction par un modèle autorégressif [JUANG et RABINER, 1985].

1.5.2 Apprentissage d'un modèle HMM

L'étape de constitution des modèles phonétiques est le point crucial de tout système RAP. L'apprentissage de ces modèles phonétiques HMM est réalisé à l'aide d'une grande base de données vocales. Une transcription phonétique est associée à chaque échantillon sonore de sorte qu'au final chaque HMM phonétique puisse être modélisé par ses représentants dans le corpus. Le nombre d'états, les transitions autorisées entre état et le symbole du phonème des modèles sont fixées et connues. Ainsi, le but de l'apprentissage est d'estimer les paramètres optimaux des HMM de chaque unité phonétique. Il nous faut donc calculer pour chaque modèle phonétique HMM :

- Les probabilités initiales π_i.
- Les probabilités de transitions a_{ij}.
- Les probabilités d'émission $b_i(o_t)$ définies par :
  - → Les vecteurs moyennes μ_{ik} (gaussienne k de l'état i).
  - → Les matrices de covariance Σ_{ik}
  - → Les poids de pondération w_{ik}.

Différentes approches d'apprentissage ont été proposées. L'approche communément utilisée s'appuie sur le critère de maximum de vraisemblance (Maximum Likelihood Estimation -MLE) estimé par l'algorithme de Baum-Welch [BAUM, 1972]. D'autres critères d'apprentissage existent, comme les critères MAP (Maximum A Posteriori) [GAUVAIN et LEE, 1994] ou MMI (Maximum Mutual Information) [BAHL et collab., 1986][NORMANDIN et collab., 1994], mais leur implémentation est plus complexe et leurs algorithmes sont plus coûteux en temps de calcul.

1.5.2.1 Estimation par maximum de vraisemblance

L'estimation par maximum de vraisemblance (Maximum Likelihood Estimation - MLE), consiste à déterminer les paramètres $\lambda = (\pi_i, a_{ij}, b_i)$ définissant un modèle HMM, qui minimisent la probabilité d'émission $P(O/\lambda)$ des observations O en terme de λ :

$$\tilde{\lambda} = \arg\max_{\lambda} P(O/\lambda) \tag{1.11}$$

Actuellement, Il n'existe pas de solution analytique directe capable de résoudre ce problème. Cependant, la méthode itérative de Baum-Welch [BAUM, 1972], qui est un cas

particulier de la méthode EM (Expectation and Maximisation) [RABINER, 1989] permet d'estimer itérativement les paramètres λ.

1.5.2.2 Algorithme de Baum-Welch

Baum a eu l'idée d'introduire une fonction auxiliaire redéfinissant le problème de recherche du modèle optimal $\tilde{\lambda}$. Ensuite il a proposé un algorithme itératif [BAUM, 1972] permettant l'estimation des nouveaux modèles.
Soit $\mathscr{B}$ une fonction auxiliaire telle que :

$$\mathscr{B}(\lambda,\lambda^{'}) = \sum_{E\in\mathscr{E}} P(O,E/\lambda)\log P(O,E/\lambda^{'}) \tag{1.12}$$

L'algorithme de Baum-Welch consiste à trouver un nouveau modèle $\lambda^{'}$ qui maximise la fonction auxiliaire $\mathscr{B}(\lambda,\lambda^{'})$.

$$\Theta(\lambda) = \arg\max_{\lambda^{'}} \mathscr{B}(\lambda,\lambda^{'}) \tag{1.13}$$

Alors :

$$\mathscr{B}(\lambda,\lambda^{'}) - \mathscr{B}(\lambda,\lambda) \leq \log P(\lambda^{'}) - \log P(\lambda) \tag{1.14}$$

Nous avons donc l'inégalité suivante :

$$P(\Theta(\lambda)) \geq P(\lambda) \tag{1.15}$$

Cet algorithme est itératif et commence par un jeu de paramètres λ_0. Ensuite, on maximise $\mathscr{B}(\lambda_0,\lambda)$ et on obtient une estimation λ_1, puis λ_2 qui maximise $\mathscr{B}(\lambda_1,\lambda)$, et ainsi de suite. Il suffit d'itérer pour obtenir des estimations toujours meilleures telles que :

$$P(\lambda_n) \geq P(\lambda_{n-1}) \geq \cdots \geq P(\lambda_2) \geq P(\lambda_1) \geq P(\lambda_0) \tag{1.16}$$

Dans le cas des modèles HMMs, $P(\lambda)$ s'écrit :

$$P(\lambda) = \sum_{E\in\mathscr{C}} \pi_{E_0} \prod_{t=1}^{T} a_{E_{t-1}E_t} b_{E_t}(o_t) \tag{1.17}$$

Où $\mathscr{C}$ représente l'ensemble des chemins possibles pour un HMM gauche-droite. Alors $\mathscr{B}(\lambda,\lambda^{'})$ peut être écrit comme la somme de trois termes (x,y,z) qui peuvent être maximisés indépendamment.

$$\mathscr{B}(\lambda,\lambda^{'}) = x(\pi_i) + y(a_{ij}) + z(b_i) \tag{1.18}$$

Concernant le premier terme, les valeurs de π_i sont constantes car elles sont fixées au moment de la construction des modèles HMMs. Pour les probabilités d'émission mono-gaussienne $\mathcal{N}(\mu_i, \Sigma_i)$ à l'état i, la ré-estimation des paramètres $(\mu_i^{'}, \Sigma_i^{'})$ du nouveau modèle $\lambda^{'}$ est décrite par les équations suivantes :

$$\begin{aligned} \mu_i^{'} &= \frac{\text{nombre de fois où on a observé } o_t \text{ à l'état } i}{\text{nombre de fois où l'on est passé par l'état } i} \\ &= \frac{\sum_{t=1}^{T} \gamma_t(i).o_t}{\sum_{t=1}^{T} \gamma_t(i)} \end{aligned} \tag{1.19}$$

Sachant que $\gamma_t(i)$ est la probabilité a posteriori d'avoir été dans l'état i du modèle λ à l'instant t connaissant l'observation O.

$$\gamma_t(i) = P(E_t = i / O, \lambda) \tag{1.20}$$

$$\Sigma_i^{'} = \frac{\sum_{t=1}^{T} \gamma_t(i)(o_t - \mu_i)(o_t - \mu_i)^{tr}}{\sum_{t=1}^{T} \gamma_t(i)} \tag{1.21}$$

Les probabilités de transitions sont ré-estimées par :

$$\begin{aligned} a_{ij}^{'} &= \frac{\text{nombre de fois où la transition de l'état } i \text{ vers l'état } j \text{ a été effectuée}}{\text{nombre de fois où l'on est passé par l'état } i} \\ &= \frac{\sum_{t=1}^{T-1} \phi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \end{aligned} \tag{1.22}$$

Où $\phi_t(i,j)$ est la probabilité d'avoir effectué la transition $i \rightarrow j$ à l'instant t connaissant l'observation O et les paramètres (μ_i, Σ_i) du modèle λ.

$$\phi_t(i,j) = P(E_t = i, E_{t+1} = j / O, \lambda) \tag{1.23}$$

La complexité de cet algorithme pour un modèle de N états est de l'ordre de $2 \times T \times N^T$, et l'ensemble des chemins $\mathscr{C}$ devient impossible à représenter. Par exemple pour 5 états et une séquence de 100 observations, cela représenterait $2 \times 100 \times 5^{100} \approx 10^{72}$ séquences (opérations !). Cependant il est possible de calculer de manière itérative γ et ϕ par deux algorithmes rapides appelés "forward-backward".

1.5.2.3 Estimation "forward-backward"

Deux variables intermédiaires sont introduites pour le calcul des inconnus γ et ϕ. La première est la variable directe $\alpha_t(i)$, définie comme la probabilité d'observer la séquence $(o_1 \dots o_t)$ et d'être à l'état i à l'instant t connaissant le modèle λ.

$$\alpha_t(i) = \mathrm{P}(o_1 o_2 \dots o_t, \mathrm{E}_t = i/\lambda) \tag{1.24}$$

La deuxième variable $\beta_t(i)$ correspond à la probabilité d'observer la séquence $o_{t+1} \dots o_\mathrm{T}$ et d'être à l'état i à l'instant t connaissant le modèle λ.

$$\beta_t(i) = \mathrm{P}(o_{t+1} \dots o_\mathrm{T}, \mathrm{E}_t = i/\lambda) \tag{1.25}$$

Par introduction de ces deux variables intermédiaires, γ et ϕ peuvent s'écrire :

$$\gamma_i(t) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{\mathrm{N}} \alpha_t(i)\beta_t(i)} \tag{1.26}$$

et

$$\phi_i(t) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{\mathrm{N}} \sum_{j=1}^{\mathrm{N}} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)} \tag{1.27}$$

α et β peuvent être calculés par récurrence sur le temps en utilisant les algorithmes "forward" et "backward" suivants :

⊛ L'algorithme directe **"forward"** :

→ Initialisation :

$$\alpha_1(i) = \begin{cases} 1, & i = 1 \\ 0, & 1 < i \leq \mathrm{N} \end{cases} \tag{1.28}$$

→ Récurrence pour t allant de 1 à T et pour j allant de 1 à N :

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{\mathrm{N}} \alpha_t(i) a_{ij} \right] b_j(o_{t+1}) \tag{1.29}$$

→ Terminaison :

$$\mathrm{P}(\mathrm{O}/\lambda) = \sum_{i=1}^{\mathrm{N}} \mathrm{P}(\mathrm{O}, \mathrm{E}_\mathrm{T} = i/\lambda) = \sum_{i=1}^{\mathrm{N}} \alpha_\mathrm{T}(i) \tag{1.30}$$

⊛ L'algorithme rétrograde **"backward"** :

→ Initialisation :

$$\beta_T(i) = \begin{cases} 1, & i = N \\ 0, & 1 \leq i < N \end{cases} \tag{1.31}$$

→ Récurrence pour t allant de T à 1 :

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \tag{1.32}$$

→ Terminaison :

$$P(O/\lambda) = \sum_{i=1}^{N} \beta_1(i) \tag{1.33}$$

Cette méthode est itéré n fois pour calculer le modèle λ_n, qui sera meilleur que le modèle λ_{n-1}. Le nombre d'itérations peut être fixé de manière arbitraire, ou bien dépendre d'un critère d'arrêt relatif à la stabilité de la convergence du modèle λ_n.

1.6 Modèle lexical

Un modèle lexical consiste à définir l'ensemble des mots qu'un système de reconnaissance de la parole peut utiliser dans la phase d'apprentissage et de test. Cet ensemble est dénommé lexique ou vocabulaire. Il est nécessaire d'associer à chaque entrée du lexique (mot) une transcription phonétique qui lui est propre. Cette modélisation est obtenue par la concaténation de phonèmes (voir la section 1.5). Une façon classique de construire le lexique consiste à extraire à partir d'un corpus textuel l'ensemble des mots les plus fréquents. Pour obtenir le dictionnaire de phonétisation, plusieurs approches sont possibles. Manuellement par des experts humains, cependant générer un lexique complet est très coûteux en ressources, et il est très difficile de couvrir la totalité des mots d'une langue. Une autre méthode possible consiste à phonétiser les mots de manière automatique [BÉCHET, 2001], en utilisant une base de règles de phonétisation pour transcrire automatiquement les graphèmes [1] en phonèmes. Le lexique doit couvrir tous les mots de la langue modélisée, et il doit tenir compte des multiples prononciations possibles d'un mot.

1. Le graphème est défini comme l'écriture associée à un phonème. Il peut être constitué d'une ou plusieurs lettres.

1.7 Modèle de langage

Les modèles de langages ont pour objectif, d'aider les Systèmes de Reconnaissance Automatique de la Parole (SRAP) dans la phase de décodage des phonèmes. Le principe est d'introduire la notion de contraintes linguistiques et les règles qui régissent le comportement de la langue modélisée. Il existe deux types de modèles de langage. Le premier est le modèle à base de grammaires formelles réalisé par des experts en linguistique, développé au début des années 1970 à partir d'automates d'états finis. De tels modèles sont encore présents dans les applications simples à vocabulaire et syntaxe limités [CHOMSKY, 1965; FU, 1971]. Le second est le modèle de langage statistique utilisant de grandes bases de données textuelles pour estimer qu'une séquence d'unités acoustiques (phonèmes, syllabes, mots, etc...) soit plus probable qu'une autre au sein de la langue modélisée. Ces modèles de langage statistiques sont privilégiés dans les systèmes RAP continue, car leur implémentation et mise en œuvre est simple et moins coûteuse en temps de calcul BAHL et collab. [1989]; JELINEK et MERCER [1980]; KATZ [1987]; KUHN et MORI [1990].

La probabilité d'une suite de k phonèmes $\mathscr{M} = (m_1 \dots m_k)$ est exprimée comme le produit des probabilités conditionnelles d'un phonème sachant tous les phonèmes précédents :

$$\mathrm{P}(\mathscr{M}) = \mathrm{P}(m_1) \prod_{i=2}^{k} \mathrm{P}(m_i / m_1 \dots m_{i-1}) \tag{1.34}$$

D'après cette théorie, la probabilité d'une séquence de plusieurs phonèmes devient rapidement proche de zéro, car aucune base de données textuelle d'apprentissage n'est suffisamment grande pour accomplir une telle modélisation. Il est donc nécessaire d'apporter des simplifications à ce modèle. Les modèles n-grammes ont ainsi été proposés [JELINEK, 1976], afin de supposer que la probabilité d'observation de la séquence de phonèmes $\mathscr{M}$ dépende uniquement des $n-1$ phonèmes précédents :

$$\mathrm{P}(\mathscr{M}) = \mathrm{P}(m_1) \prod_{i=2}^{n-1} \mathrm{P}(m_i / m_1 \dots m_{i-1}) \prod_{i=n}^{k} \mathrm{P}(m_i / m_{i-n+1} \dots m_{i-1}) \tag{1.35}$$

Lorsque n vaut 2 ou 3, on parlera respectivement de modèles bigrammes (un phonème dépend du phonème qui le précède) et trigrammes (un phonème dépend des deux phonèmes qui le précèdent) [JELINEK et MERCER, 1980]. Ces deux modèles sont les plus utilisés dans les systèmes de reconnaissance de la parole continue en fonction de la quantité de données exploités.

Dans une modélisation trigramme l'équation précédente peut être simplifiée par :

$$P(\mathscr{M}) = P(m_1)P(m_2/m_1)\prod_{i=3}^{k} P(m_i/m_{i-2}m_{i-1}) \tag{1.36}$$

1.7.1 Estimation des modèles de langage

Le critère de maximum de vraisemblance (Maximum Likelihood – ML) est utilisé pour estimer les probabilités d'un modèle de langage n-grammes.

$$P(m_i/m_{i-n+1}\dots m_{i-1}) = \frac{\mathscr{O}(m_{i-n+1}\dots m_{i-1}m_i)}{\mathscr{O}(m_{i-n+1}\dots m_{i-1})} \tag{1.37}$$

Où $\mathscr{O}(m_{i-n+1}\dots m_{i-1})$ représente le nombre d'occurrences de la séquence de phonèmes $(m_{i-n+1}\dots m_{i-1})$ dans le corpus textuel d'apprentissage. Certainement les séquences de phonèmes n'apparaissent pas toutes dans la partie apprentissage d'une base de données et par conséquent, une probabilité nulle ne peut être attribuée. La technique de lissage permet de remédier à ce problème, en combinant les modèles (trigramme, bigramme et unigramme).

1.7.2 Évaluation du modèle de langage

Le modèle de langage permet de guider le décodage pour améliorer la performance des SRAP. La perplexité (PPL) est une méthode rapide pour évaluer la capacité de prédiction des modèles de langage [JELINEK et collab., 1977]. Plus le modèle de langage est meilleur et performant, plus la valeur de perplexité est faible.

$$\log(\text{PPL}) = -\frac{1}{n}\sum_{i=1}^{n} \log P(m_i/m_1\dots m_{i-1}) \tag{1.38}$$

1.8 Décodage de la parole continue

Le décodage des phrases prononcées est un processus délicat, car en parole continue, la segmentation de ces phrases de test en phonèmes ainsi que le nombre de phonèmes que comporte chaque phrase ne sont pas connus. Le but du décodage alors est de déduire la séquence d'états qui a généré les observations données. En effet, nous pouvons facilement trouver la suite de phonèmes la plus probable qui correspond aux pa-

ramètres observés à partir de cette séquence d'états. Cette tâche est accomplie grâce à l'algorithme de recherche Viterbi [VITERBI, 1967] à l'aide des probabilités générées par les modèles phonétiques HMM et les probabilités du modèle de langage. L'exploration de l'algorithme de recherche Viterbi (appelé aussi Beam Search) est effectuée à chaque étape sur les meilleurs chemins. Un graphe d'états (voir figure 1.4) est mis à jour en permanence pour représenter l'ensemble des hypothèses de transcription et ainsi trouver le chemin optimal qui correspond à la séquence de phonèmes prononcées.

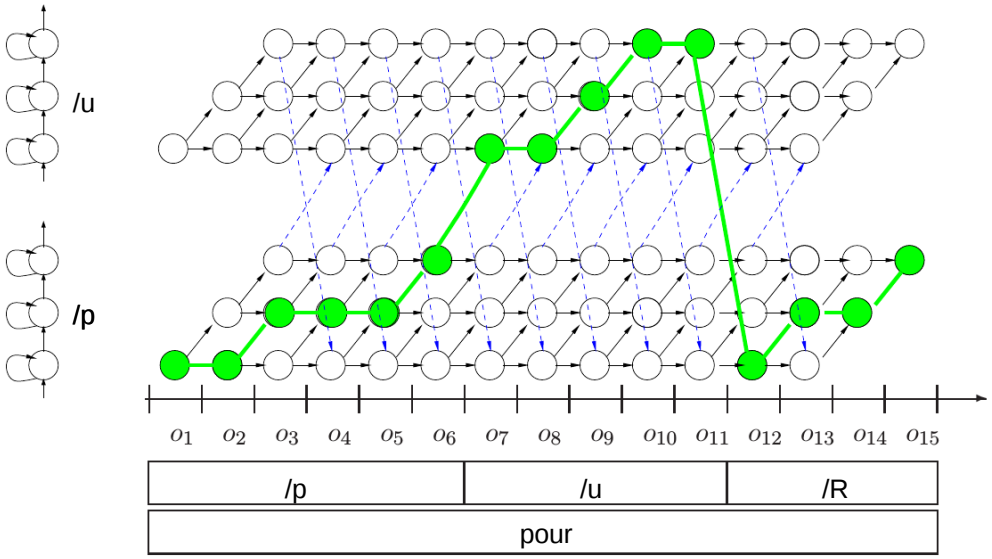

FIGURE 1.4: *Décodage Viterbi : Pour cet exemple la meilleur hypothèse correspond à la succession de phonèmes /p /u /R qui est la transcription phonétique du mot "pour".*

L'algorithme de recherche Viterbi est un algorithme de programmation dynamique similaire à l'algorithme "forward". Cet algorithme peut être décrit par les étapes suivantes :

⊛ **Algorithme Viterbi** :

→ Initialisation :

$$\delta_1(i) = \pi_i b_i(o_1) \quad \text{et} \quad \psi_1(i) = 0 \tag{1.39}$$

→ Récurrence : pour t allant de 1 à T (nombre d'observations)

pour j allant de 1 à N (nombre d'états)

$$\delta_t(j) = \max_{1 \le i \le N} \left(\left[\delta_{t-1}(i) a_{ij} \right] b_j(o_t) \right) \tag{1.40}$$

$$\psi_t(j) = \arg\max_{1 \le i \le N} \left(\left[\delta_{t-1}(i) a_{ij} \right] \right) \tag{1.41}$$

→ Terminaison :

$$\hat{P} = \max_{1 \le i \le N} \delta_T(i) \tag{1.42}$$

$$\hat{E}_T = \arg\max_{1 \le i \le N} \delta_T(i) \tag{1.43}$$

Où $\delta_t(i)$ correspond à la vraisemblance du meilleur chemin qui finit à l'état i au temps T. $\psi_t(i)$ correspond à un tableau de traces pour stocker l'état i (qui précède l'état actuel j) utilisé pour calculer le maximum de $\delta_t(i)$. Le principe revient à construire de façon itérative la meilleure séquence d'états à partir de ce tableau de traces. Pour retrouver le chemin optimal et la chaîne de phonèmes, il faut retourner en arrière à partir de l'état qui maximise $\delta_T(i)$.

la meilleure séquence d'états est :

$$\hat{E}_t = \psi_{t+1}(\hat{E}_{t+1}) \quad \text{pour} \quad t = T-1, T-2, \ldots, 1 \tag{1.44}$$

La complexité de cet algorithme est de $N^2 \times T \times U + U^2 \times T$, avec U représente le nombre de modèles HMM phonétiques. Il est toujours possible d'effectuer les calculs en temps réel dans le cas d'utilisation d'un ensemble réduit de modèles phonétiques (monophones). En revanche, l'utilisation des modèles phonétiques dépendants du contexte (triphones) nécessite plus de temps de calcul dans la reconnaissance.

1.8.1 Évaluation du module de décodage

Le décodage de la parole continue fournit comme sortie, des séquences de phonèmes. Chaque séquence permet de représenter une phrase reconnue parmi les phrases de test. Deux mesures sont couramment utilisées pour évaluer le décodage de phonèmes. Il y a le taux d'erreur de phonèmes (Phone Error rate - PER), ou la mesure de performance connexe : taux de reconnaissance de phonèmes (Accuracy). Ces mesures sont calculées respectivement par les équations 1.45 et 1.46.

$$\text{PER} = \frac{I + O + S}{N_T} \tag{1.45}$$

$$\text{Accuracy} = \frac{N_T - (I + O + S)}{N_T} \tag{1.46}$$

Où N_T représente le nombre total d'étiquettes contenu dans l'énoncé de référence de test et S,I et O correspondent respectivement aux erreurs de Substitution, d'Insertion et d'Omission. Ces erreurs sont calculées par un algorithme de programmation dynamique DTW (Dynamic Time Warping) [VINTSYNK, 1968][SAKOE et CHIBA, 1971] qui compare la chaine de phonèmes correcte (de référence) et la chaine de phonèmes reconnue (de test). La performance d'un système RAP continue peut être calculée à l'aide d'une autre mesure supplémentaire. Cette mesure représente le taux de phonèmes correct (Correct). Elle est similaire à la précision (Accuracy), la seule différence est que les insertions (I) ne sont pas considérées comme des erreurs, donc sont ignorées.

$$\text{Correct} = \frac{N_T - (O + S)}{N_T} \tag{1.47}$$

1.9 Conclusion

L'objectif d'un système de reconnaissance automatique de la parole continue, est de reconnaître la séquence de phonèmes contenue dans un signal de la parole à l'aide d'un dispositif informatique. Malgré des efforts considérables et quelques avancées spectaculaires, la capacité d'une machine à reconnaître la parole est encore loin d'égaler celle de l'être humain. En effet, le signal vocal est très complexe à analyser car il ne transporte pas seulement le message linguistique émis par un locuteur, mais aussi un ensemble d'informations sur ce locuteur. Plusieurs facteurs sont à l'origine de cette complexité, en particulier la redondance, la continuité et les effets de coarticulation, ainsi que l'ample variabilité intra et inter-locuteurs. Toutes ces caractéristiques rendent très difficile la tâche d'un système RAP. Pour pallier ces problèmes, plusieurs approches ont été proposées. Cependant, la plupart des SRAP continues sont construits en utilisant des modèles statistiques (modèles de Markov cachés HMM). En effet, le temps qui a été consacré à leur mise au point est nettement supérieur à celui qui a été dédié aux nouvelles solutions. Ainsi, nous avons décrit clairement dans ce chapitre les bases théoriques et les différents concepts qui sous-tendent l'élaboration des SRAP basés sur les modèles HMM.

Chapitre 2

Reconnaissance automatique de la parole laryngée

« *La patience est la clé du bien-être.* »

Mohammed

La tradition musulmane - VIIe siècle.

2.1 Introduction

Notre objectif dans un premier temps est de construire un système de reconnaissance de la parole continue indépendant du locuteur. Nous avons réussi à créer notre propre système nommé SPIRIT [LACHHAB et collab., 2012], mis en œuvre à partir de modèles HMMs avec des hypothèses d'apprentissage et de test très simples et non coûteuses en temps de calcul. En outre, ce système modélise la durée d'émission des modèles phonétiques pour améliorer le taux de décodage de la parole. Ultérieurement, nous avons implémenté deux autres systèmes RAP à l'aide de la plate-forme HTK (Hidden Markov Model Toolkit [YOUNG et collab., 2006], qui intègrent plusieurs méthodes complexes par exemple les algorithmes : Baum-Welch, Viterbi et DTW permettant une meilleure estimation des paramètres HMM avec un décodage rapide de la parole. Le deuxième système est monophone construit en utilisant des modèles phonétiques indépendants du contexte. Plusieurs expériences ont été effectuées avec ce système, comme par exemple la variation du nombre de gaussiennes utilisées dans chaque état et du nombre de coefficients des vecteurs acoustiques. Nous avons aussi examiné l'évolution des taux de décodage après l'utilisation d'un modèle de langage bigramme. Un troisième système plus performant a ensuite été développé à partir du système monophone en utilisant des modèles phonétiques dépendants du contexte (triphones). De plus, les performances du système triphone ont été améliorées par la transformation HLDA des vecteurs acoustiques pour réduire leur dimension dans un espace restreint ayant de bonnes propriétés discriminantes. Les résultats expérimentaux démontrent que nos systèmes fournissent des améliorations significatives du taux de reconnaissance phonétique (Accuracy) sur la partie noyau de la partie test du corpus TIMIT.

2.2 Base de données TIMIT

Nous avons choisi d'évaluer nos systèmes de reconnaissance automatique de la parole laryngée avec la base de données acoustiques TIMIT [GAROFOLO et collab., 1993] pour plusieurs raisons. Tout d'abord, parce qu'elle est une base de référence communément utilisée par les chercheurs pour comparer leurs résultats. Deuxièmement, parce qu'elle est fournie avec une segmentation phonétique manuelle, qui simplifie l'apprentissage des modèles phonétiques d'un système RAP continue. De plus, les accents couramment

utilisés dans diverses régions des États-Unis (voir le tableau 2.1) sont convenablement illustrés dans cette base de données TIMIT.

Dialecte	Régions	Homme	Femme	Total
1	New England	31 (63%)	18 (25%)	49 (8%)
2	Northern	71 (70%)	31 (30%)	102 (16%)
3	North Midland	79 (67%)	23 (23%)	102 (16%)
4	South Midland	69 (69%)	31 (31%)	100 (16%)
5	Southern	62 (63%)	36 (37%)	98 (16%)
6	New York City	30 (65%)	16 (35%)	46 (7%)
7	Western	74 (74%)	26 (26%)	100 (16%)
8	Army Brat	22 (67%)	11 (33%)	33 (5%)

TABLEAU 2.1: *Distribution des 8 dialectes de la base de données TIMIT*

2.2.1 Description de la base TIMIT

TIMIT est un corpus de parole dédié à la reconnaissance de la parole continue indépendante du locuteur. Dans cette base de données, 630 locuteurs américains répartis sur 8 dialectes régionaux (“dr1” à “dr8”) ont participé à la procédure d’enregistrement sonores des phrases. Chaque locuteur a prononcé 10 phrases différentes choisis comme suit :

- ⊛ 2 phrases (identifiées “sa1.wav” et “sa2.wav”) dites de calibration, pour élucider les diversités dialectiques régionales.
- ⊛ 5 phrases phonétiquement équilibrées (identifiés “sx3.wav” à “sx452.wav”).
- ⊛ 3 phrases sont choisies pour illustrer la variation phonétique contextuelle (identifiées “si453.wav” à “si2342.wav”). L’enregistrement sonore des phrases c’est déroulé dans de bonnes conditions (le signal sonore est échantillonné à 16KHz avec 16 bits de codage pour chaque échantillon). Ce corpus, possède un vocabulaire total de 6100 mots. La répartition globale des locuteurs par genre est de 438 hommes et 192 femmes représentée comme suite :
  - → Dans la partie apprentissage : 326 hommes et 136 femmes.
  - → Dans la partie test : 112 hommes et 56 femmes.

Les locuteurs hommes sont identifiés par la lettre “m” tandis que les femmes sont identifiées par la lettre “f”. Un sous-ensemble de test, appelé noyau de test (en anglais

Core Test), ne contient que 192 phrases prononcées par 24 locuteurs (2 hommes et une femme pour chacun des 8 dialectes). Le core test comporte 7215 segments phonétiques (les phrases de calibration sont exclues). Sa taille réduite par rapport à la partie test complète (1344 phrases), permet de multiplier les expériences tout en préservant un calcul réaliste des taux de reconnaissance réels. Chaque enregistrement sonore est fourni avec 3 autres fichiers portant le même nom avec les extensions suivantes :

- ".txt" : transcription textuelle de la phrase prononcée suivi du nombre d'échantillons totale de l'enregistrement.
- ".phn" : segmentation phonétique manuelle avec le nombre d'échantillons de chaque phonème.
- ".wrd" : transcription orthographique en mots avec le nombre d'échantillons de chaque mot.

Les fichiers sons ".wav" sont échantillonnés a 16 Khz, donc la durée en secondes correspond au nombre d'échantillons divisé par 16000. Cette base de données, utilise un étiquetage de 61 phonèmes différents. La liste de tous ces phonèmes est représentée dans le tableau 2.2, avec leur équivalent dans l'Alphabet Phonétique International (API) suivi d'un exemple de composition dans un mot anglais.

2.2.2 Étiquetage Kai-Fu Lee (KFL)

L'étiquetage d'origine en 61 phonèmes est jugé trop détaillé pour l'apprentissage des modèles phonétiques. [Lee et Hon, 1989] ont proposé de réduire le nombre de classes phonétiques à 39 seulement au lieu de 61 par le regroupement des allophones. Cette étiquetage a été ensuite utilisé dans la plupart des travaux de recherches. Ce regroupement est réalisé en deux phases :

- Avant l'apprentissage, les 61 phonèmes d'origine sont réduits en 48 classes phonétiques par fusion d'allophones (ax/ax-h, er/axr, hh/hv, m/em, ng/eng, n/nx, ux/uw), regroupement des silences dans une nouvelle étiquette 'sil' pour les silences h#/pau, les occlusives précédant un arrêt voisé (bcl/dcl/gcl) sont remplacées par une occlusive voisée 'vcl' et les occlusives sourdes (pcl/tcl/kcl) sont remplacées par une occlusive non voisée 'cl'. Enfin l'étiquette 'q' qui ne correspond pas toujours à une occlusive est supprimée.

⊛ Lors du calcul des taux de reconnaissance (test), les confusions (aa/ao, ax/ah, ih/ix, l/el, n/en, sil/epi/cl/vcl, sh/zh) sont permises conduisant à un regroupement en 39 classes phonétiques.

TIMIT	API	Exemple	TIMIT	API	Exemple	TIMIT	API	Exemple
		Occlusives :			Nasales :			Voyelles :
pcl p	p	**p**ea	m	m	**m**om	iy	i	b**ee**t
tcl t	t	**t**ea	em	ɱ	bott**om**	ih	ɪ	b**i**t
kcl k	k	**k**ey	n	n	**n**oo**n**	ix	ɨ	deb**i**t
bcl b	p	**b**ee	nx	ɾ̃	wi**nn**er	eh	ɛ	b**e**t
dcl d	p	**d**ay	en	ɳ	butt**on**	ae	æ	b**a**t
gcl g	p	**g**ay	ng	ŋ	s**ing**	aa	ɑ	b**o**tt
dx	ɾ	mu**dd**y	eng	ŋ̍	wash**ing**ton	ao	ɔ	b**ough**t
q	ʔ	ba**t**			Liquides :	uh	ʊ	b**oo**k
		Affriquées :	l	l	**l**ay	uw	u	b**oo**t
dcl jh	dʒ	**j**oke	el	ɬ	bott**le**	ux	ü	t**oo**t
tcl ch	tʃ	**ch**oke	r	r	**r**ay	ax	ə	**a**bout
		Fricatives :			Semi-voyelles :	ax-h	ə̊	s**u**spect
f	f	**f**in	w	w	**w**ay	ah	ʌ	b**u**t
th	θ	**th**in	y	j	**y**acht	er	ɝ	b**ir**d
s	s	**s**ea			Fricatives glottale :	axr	ɚ	butt**er**
sh	ʃ	**sh**e	hh	h	**h**ay			Diphtongues :
v	v	**v**an	hv	ɦ	a**h**ead	ey	e	b**ai**t
dh	ð	**th**en			Silences :	ay	α^y	b**ite**
z	z	**z**one	h#			oy	$\ɔ^y$	b**oy**
zh	ʒ	a**z**ure	pau	api		aw	α^w	b**ou**t
						ow	o	b**oa**t

TABLEAU 2.2: *Etiquetage de TIMIT, code API correspondant et exemple de mot anglais contenant le phonème.*

Le tableau 2.3 présente des statistiques sur les 48 phonèmes d'apprentissage. Pour chaque classe phonétique, nous donnons le nombre de représentants ou d'échantillons ainsi que sa durée moyenne. Le regroupement des allophones est mentionné par virgule, tandis que les confusions autorisées entre phonèmes dans la phase de reconnaissance ont été encadrées.

Etiquette	Nombre	Durée (ms)	Etiquette	Nombre	Durée (ms)
Occlusives :			Semi-voyelles :		
b	2181	17	w	2216	60
d	2432	24	y	995	54
g	1191	27	Fricative glottale :		
p	2588	44	hh,hv	1660	67
t	3948	49	Voyelles :		
k	3794	52	iy	4626	95
dx	1864	29	ih	4248	78
Affriquées :			ix	7370	51
jh	1013	61	eh	3277	93
ch	820	86	ae	2292	136
Fricatives :			aa	2256	123
f	2215	103	ao	1865	123
th	745	92	uh	500	76
s	6176	113	uw,wx	1952	100
sh	1317	118	ax,ax-h	3892	47
zh	149	81	ah	2266	89
v	1994	60	er,axr	4138	95
dh	2376	36	Diphtongues :		
z	3682	84	ey	2271	127
Nasales :			ay	1934	155
m,em	3566	65	oy	304	168
n,nx	6896	52	aw	728	161
en	630	78	ow	1653	128
ng,eng	1220	61	Silences :		
Liquides :			sil=(h#,pau)	8283	191
l	4425	61	cl=(pcl,tcl,kcl)	12518	58
el	951	90	vcl=(bcl,dcl,gcl)	7219	54
r	4681	56	epi	908	42

TABLEAU 2.3: *Statistiques sur le nombre d'échantillons et la durée moyenne des 48 classes phonétiques (les confusions autorisées dans la phase de décodage sont encadrées).*

2.3 Système SPIRIT

Dans cette section, nous décrivons notre propre système de reconnaissance automatique de la parole laryngée nommé SPIRIT [LACHHAB et collab., 2012]. Ce système s'appuie sur les algorithmes d'apprentissage conçus au sein de l'équipe Parole de Nancy, sur la reconnaissance de phonèmes isolés en utilisant la base de données TIMIT. Nous avons réussi à adapter et appliquer ces méthodes à la reconnaissance de phonèmes connectés indépendante du locuteur. Les modèles phonétiques indépendants du contextes sont estimés directement à partir des données au lieu d'utiliser la procédure classique Baum-Welch. Une modélisation de la durée d'émission des modèles phonétique HMM basée sur une distribution gaussienne a été proposée pour améliorer le taux de décodage de la parole de ce système.

2.3.1 Prétraitement des données

Il est absolument primordial de transformer le signal de la parole en vecteurs acoustiques. Nous utilisons pour notre système SPIRIT, les vecteurs MFCC. Tout d'abord le signal est échantillonné à 16 Khz et pré-accentué avec un facteur de 0.96. Chaque trame est multipliée par une fenêtre de Hamming de 32 ms décalée toute les 10 ms afin de maintenir la continuité des premiers et derniers points. Chaque vecteur comporte 11 coefficients cepstraux statiques, calculés en utilisant un banc de 26 filtres en échelle Mel. Le logarithme de l'énergie de la trame est ajouté à ces 11 coefficients pour former des vecteurs de 12 coefficients. Les dérivées d'ordre 1 et 2 (Δ et $\Delta\Delta$) sont calculées par notre propre formule suivante :

$$\Delta x_t(c) = x_t(c+1) - x_t(c-1) \tag{2.1}$$

Où $x_t(c)$ représente le coefficient c du vecteur statique de la trame t et $\Delta x_t(c)$ son coefficient différentiel d'ordre 1 correspondant. La dérivée d'ordre 2 et les dérivées de l'énergie sont calculés de la même façon. Donc nous travaillons avec des vecteurs MFCC de dimension $d = 36$ (11 MFCC; E; 11 ΔMFCC; ΔE; 11 $\Delta\Delta$MFCC; $\Delta\Delta$E).

2.3.2 Apprentissage des modèles phonétiques

Notre système SPIRIT est basé sur des modèles phonétiques HMMs indépendants du contexte. Pour faire l'apprentissage et le décodage, 39 modèles phonétiques issus de la

classification de Kai Fu lee (voir la section 2.2.2) ont été utilisés. Chaque phonème correspond à un HMM gauche-droit composé de 5 états (mais seulement 3 entre eux sont émetteurs). Les probabilités d'émissions sont estimées en distribution continue par une somme pondérée de G gaussiennes multivariées (GMM). Chaque gaussienne est représentée par un vecteur moyen (centroïde) μ et une matrice de covariance Σ. Les centroïdes μ_{ik} sont estimés initialement en utilisant l'algorithme de quantification vectorielle LBG[LINDE et collab., 1980] appliqué sur les vecteurs associés à l'état i. Chaque centroïde k de l'état i (μ_{ik}) est calculé par une moyenne de ses vecteurs cepstraux associés x_{ik}^n où x_{ik}^n est le $n^{ème}$ vecteur de la classe k de l'état i.

$$\mu_{ik} = \frac{1}{N_{ik}} \sum_{n=1}^{N_{ik}} x_{ik}^n \tag{2.2}$$

Où N_{ik} représente le nombre des vecteurs associés au centroïde k de l'état i. Les matrices de covariances Σ_{ik} sont calculés aussi statistiquement à partir des données en utilisant la formule suivante :

$$\Sigma_{ik} = \frac{1}{N_{ik}} \sum_{n=1}^{N_{ik}} (x_{ik}^n - \mu_{ik})(x_{ik}^n - \mu_{ik})' \tag{2.3}$$

Le poids de pondération w_{ik} de la gaussienne k est estimé par la formule suivante :

$$w_{ik} = \frac{N_{ik}}{N_i} \tag{2.4}$$

N_i et N_{ik} correspondent respectivement au nombre de vecteurs cepstraux associés à l'état i et au nombre de vecteurs cepstraux associés à la gaussienne k de l'état i.

Le choix du nombre de gaussiennes utilisé dans chaque état est très important parce qu'il peut influencer le taux de reconnaissance. Un mauvais apprentissage peut être observé lors de l'utilisation d'un nombre trop élevé de gaussiennes vu la quantité de données d'apprentissage disponible. Pour cette raison, nous commençons par 16 gaussiennes dans chaque état. Ce nombre de gaussiennes est optimisé en fonction du nombre de vecteurs MFCC associés à chaque état : si ce dernier est inférieur à la dimension d des vecteurs, alors la gaussienne associée est supprimée. Les vecteurs associés à cette gaussienne supprimée sont redistribués sur les plus proches centroïdes.

Nous estimons les probabilités de transition entre états en utilisant la loi géométrique. Soit $\mathscr{X}$ une variable aléatoire donnant le nombre de fois que l'état à été visité. Si on considère les événements $\mathscr{R}_j$ "Rester j fois dans le même état" et $\mathscr{M}_j$ "Passer à l'état suivant au moment j". Alors l'événement $[\mathscr{X} = l]$ peut être formulé par :

$$[\mathscr{X} = l] = \underbrace{\mathrm{R}_1 \cap \mathrm{R}_2 \cap \ldots \cap \mathrm{R}_{l-1}}_{\mathscr{R}_j} \cap \underbrace{\mathrm{M}_l}_{\mathscr{M}_j} \tag{2.5}$$

les événements sont indépendants, donc la probabilité de distribution de $\mathscr{X}$ peut être calculé par la formule suivante :

$$p(\mathscr{X} = l) = p_r^{l-1}.p_m \tag{2.6}$$

Où p_r est la probabilité de rester dans le même état et $p_m = 1 - p_r$ est la probabilité de passer à l'état suivant.

L'espérance de cette variable $\mathscr{X}$ est donnée par :

$$\mathrm{E}[\mathscr{X}] = \sum_{l=1}^{+\infty} l.p_r^{l-1}(1-p_r) = \frac{1}{1-p_r} \tag{2.7}$$

Donc

$$p_r = \frac{\mathrm{E}[\mathscr{X}] - 1}{\mathrm{E}[\mathscr{X}]} \tag{2.8}$$

l'espérance $\mathrm{E}[\mathscr{X}]$ est calculé directement à partir des données par la formule suivante :

$$\mathrm{E}[\mathscr{X}] = \frac{\mathrm{N}_{ip}}{\mathrm{N}_p} \tag{2.9}$$

Où N_{ip} représente le nombre de vecteurs associés à l'état i du phonème p et N_p correspond au nombre total d'échantillons du phonème p.

L'algorithme de Viterbi a été appliqué sur les vecteurs MFCC de chaque phrase pour raffiner l'apprentissage des modèles. Cette algorithme est itéré au maximum 20 fois ou jusqu'à avoir une stabilité au niveau des chemins retournés par ce processus de Viterbi.

2.3.3 Décodage de la parole

Notre système SPIRIT est un système de reconnaissance automatique de la parole continue. Le décodage est effectué par l'algorithme classique Viterbi en utilisant les 39

modèles phonétiques déjà appris. La recherche de la meilleure chaîne de phonèmes qui a généré les vecteurs en entrée du SRAP est améliorée par l'inclusion d'un modèle de langage bigramme et un modèle de durée. Le modèle de langage bigramme correspond à un tableau à deux dimensions contenant la probabilité d'occurrence de deux phonèmes successifs. Notre modèle de durée suit une loi normale $\mathcal{N}(\mu, \sigma^2)$ estimée pour chaque phonème selon le nombre de vecteurs contenus dans chaque modèle phonétique HMM au cours du décodage.

$$\mu = \frac{N_{vp}}{N_{ep}} \tag{2.10}$$

$$\sigma^2 = \frac{N_{vp}^2}{N_{ep}} - \mu^2 \tag{2.11}$$

Où N_{vp} représente le nombre de vecteurs du phonème p et N_{ep} correspond au nombre d'échantillons de ce phonème. La probabilité de la durée est intégrée au cours du décodage au niveau de la transition inter-états.

2.3.4 Expériences et résultats

Notre système SPIRIT a été évalué sur la base de données TIMIT. L'apprentissage des 39 modèles phonétiques HMM est effectué sur 3696 phrases, tandis que le décodage est réalisé sur la totalité de la partie test du corpus TIMIT. Cette partie de test contient 1344 phrases composées de 50754 phonèmes.

Les taux de reconnaissance sont représentés dans le tableau 2.4, soit en utilisant un modèle de langage bigramme seul ou avec l'ajout du modèle de durée.

39 monophones	**Bigramme**	**Bigramme+Durée**
Substitution	17.61% (8938)	17.25% (8756)
Omission	10.46% (5310)	11.69% (5932)
Insertion	7.11% (3607)	5.81% (2951)
Correct	71.93% (36506)	71.06% (36066)
Accuracy	64.82% (32899)	65.25% (33115)

TABLEAU 2.4: *L'influence d'un modèle de durée sur le taux de reconnaissance phonétique.*

D'après le tableau ci-dessus, nous remarquons que le modèle de durée ajouté dans le processus de décodage, permet de réduire le nombre des erreurs d'insertion et par conséquence d'améliorer le taux de reconnaissance phonétique (Accuracy).

2.4 Plate-forme HTK

En 1995, S.J. Young et son équipe ont développé à l'université de Cambridge la plate-forme HTK (Hidden Markov Model Toolkit). Cette boîte à outils open source, se compose d'un ensemble de module permettent de faciliter la mise en œuvre d'un système RAP continue à base des HMM [YOUNG et collab., 2006]. Nous avons donc choisi de construire notre système de référence pour la reconnaissance de la parole laryngée à partir de cette plate-forme HTK pour plusieurs raisons :

- D'abord, parce que cette boite à outils intègre les différents algorithmes classiques d'apprentissage et de décodage utilisés dans les système RAP (Baum-Welch, Viterbi, DTW, etc.).
- Deuxièmement, l'ensemble des outils est écrit en langage C, et la documentation détaille leur utilisation et les principes de leur implémentation : ceci permet d'intégrer de manière efficace les modifications souhaitées.
- En plus, HTK est largement répondu dans le monde de la recherche : celui-ci permet d'évaluer (ou comparer) de manière plus précise les résultats.

Toutes les fonctionnalités d'HTK sont définies par des modules assurant l'utilisation des outils de base (voir tableau 2.5). Ces outils permettent d'analyser le signal de la parole, de manipuler les transcriptions des mots et des phonèmes, de définir des modèles acoustiques et de langage, de faire l'apprentissage et l'adaptation de ces modèles, d'aligner et décoder la parole continue etc. Les options d'utilisation des outils sont transmises en argument sur la ligne de commande. Il est donc facile d'automatiser le processus d'extraction des paramètres acoustiques, d'apprentissage et de décodage avec des scripts écrits dans un langage de commande (par exemple dans notre cas en C-Shell sous Ubuntu (UNIX)).

	Librairies		**Outils de base**
HShell	Interface système d'exploitation	HLEd	Edition des fichiers d'étiquettes
HMath	Procédures mathématiques	HHEd	Edition des modèles
HSigP	Procédures de traitement du signal	HCopy	Calcul des paramètres du signal
HDBase	Stockage en mémoire des paramètres	HBuild	Formatage des modèles de langage
HSpIO	Transformations du signal	HCompV	Calcul des moyennes et variances
HAudio	Acquisition du signal	HDMan	Manipulation des dictionnaires
HWave	Gestion du signal	HParse	Génération du graphe de décodage
HParm	Calcul des paramètres d'exploitation	HQuant	QV pour modèles discrets
HVQ	Gestion de la QV	HSGen	Génération aléatoire de phrases test
HLabel	Gestion des fichiers d'étiquettes	HSmooth	Lissage des paramètres des modèles
HTrain	Gestion de l'apprentissage	HInit	Initialisation d'un modèle
HLM	Gestion des modèles de langage	HRest	Réestimation d'un modèle
HNet	Gestion des réseaux	HERest	Réestimation des modèles enchaînés
HDict	Gestion des dictionnaires	HVite	Décodage en parole continue
HParse	Lecture du réseau syntaxique	HResults	Résultats du décodage
HGraf	Affichage graphique	HList	Affichage des fichiers de données
		HLStats	Calcul de statistiques
		HSLab	Affichage du signal et des étiquettes

TABLEAU 2.5: *Librairies et outils de base d'HTK.*

2.5 Système de reconnaissance monophone

Les séquences de mots sont modélisées par un ensemble d'unités acoustiques, fréquemment les phonèmes. Pour le développement d'un système de reconnaissance monophone (indépendant du contexte), chaque phonème doit être modélisé par un seul HMM gauche-droite à cinq états (voir la figure 1.3). L'état initial et l'état final ont pour objectif de servir uniquement à la connexion des modèles en parole continue sans émettre d'observation. Pour modéliser les 48 phonèmes du regroupement de Kai-Fu Lee [LEE et HON, 1989], nous avons besoins de 48 HMMs et le nombre total d'états est alors 144 seulement. Les probabilités d'émission sont calculées par une somme pondérée de G gaussiennes multivariées(GMM), caractérisées par leur vecteur moyen et leur matrice de covariance. L'apprentissage des modèles phonétiques en utilisant une matrice de covariance non diagonale est très coûteux en mémoire et temps de calcul par rapport au cas d'utilisation d'une matrice de covariance diagonale. En effet, une matrice de covariance non diagonale contient un nombre de paramètres considérablement élevé. Pour cette raison,

nous avons choisi un apprentissage à l'aide des matrices de covariance diagonales.

2.5.1 Prétraitement des données

Le système de reconnaissance monophone utilise les coefficients MFCC et l'énergie, ainsi que les coefficients différentiels de ces paramètres (voir la section 1.4.1 et 1.4.2). Le module HCopy de la plate-forme HTK permet de transformer les enregistrements TIMIT (.wav) en vecteurs MFCC (.mfcc).

La configuration utilisée est la suivante :

→ Signal échantillonné à 16 Khz.

→ Pré-accentué avec un facteur de 0.97.

→ Fenêtre de Hamming de 25 ms.

→ Pas de décalage entre deux trames successives : 10 ms.

→ Banc de 26 filtres en échelle Mel.

→ Conservation des 12 premiers coefficients cepstraux et concaténation avec le logarithme de l'énergie de la trame pour former un vecteur de 13 coefficients statiques.

→ Ajout des coefficients différentiels dits "dynamiques" d'ordre 1 et 2 (Δ et $\Delta\Delta$).

Pour ce système de référence, 39 coefficients au total sont calculés pour chaque trame. Ce nombre de coefficient ($d = 39$), représente le nombre référence de la dimensionnalité utilisée dans la plupart des systèmes RAP continue.

2.5.2 Apprentissage des modèles monophones

Les 48 modèles HMMs monophones de la classification de Kai Fu Lee (voir la section 2.2.2) représentant le vocabulaire phonétique de la base TIMIT doivent d'abord être initialisés. Cette procédure est effectuée par l'outil HInit en utilisant l'algorithme itératif des "k-moyennes segmentales" basée sur l'algorithme de Viterbi. Cette étape nécessite l'étiquetage des phrases d'apprentissage en fonction des unités acoustiques modélisées (48 phonèmes indépendants du contexte). L'outil HLEd permet de modifier l'étiquetage pour remplacer, fusionner ou supprimer un ou plusieurs segments phonétiques. L'estimation des probabilités d'émission des observations (vecteurs MFCC) et des probabilités de transition entre états est calculée en utilisant l'algorithme de Baum-Welch à l'aide de l'outil

HRest. L'étape finale de l'apprentissage consiste à ré-estimer simultanément l'ensemble des modèles sur la parole continue grâce à l'outil HERest.

Nous pouvons améliorer les modèles monophones en augmentant le nombre de gaussiennes permettant d'estimer la probabilité d'émission d'un vecteur dans un état. Cependant il est essentiel de choisir le nombre nécessaire de gaussiennes attribuées à chaque état, en faisant une meilleure adaptation entre une adéquate modélisation des HMM monophones et le nombre limité de données d'apprentissage. Le problème qui se pose alors est de trouver le nombre de composantes qui est le mieux adapté aux données disponibles. Un nombre élevé de gaussiennes, conduit à un mauvais apprentissage, parce que les données d'apprentissage ont un nombre limité d'échantillons pour chaque phonème. De plus, l'estimation des différents paramètres optimaux des modèles HMM monophones sera très coûteuse en mémoire et aussi en temps de calcul. Pour optimiser le nombre de gaussiennes utilisées dans chaque état [JOUVET et collab., 1991] proposent une augmentation successive du nombre de gaussiennes suivie de fusions des gaussiennes les plus proches. Cette procédure permet de supprimer les gaussiennes qui sont estimées avec un nombre de vecteurs trop faible.

Dans notre système, le nombre de gaussiennes peut être choisi soit dans la configuration des modèles ou augmenté de manière itérative par l'intermédiaire de l'outil HHEd. Il faut noter, que dans le deuxième cas les modèles HMMs monophones doivent être réestimés après chaque incrémentation itérative du nombre de gaussiennes.

L'augmentation des gaussiennes se fait par clonage et perturbation. Par exemple dans l'état i d'un modèle HMM, la probabilité d'émission des observations O est calculée par un mélange de G gaussiennes dont les paramètres w_i, μ_i, Σ_i ont été estimés par l'algorithme Baum-Welch :

$$b_i(\mathrm{O}) = \sum_{k=1}^{\mathrm{G}} w_k \mathcal{N}(\mathrm{O}, \mu_k, \Sigma_k) \tag{2.12}$$

Alors pour doubler le nombre de gaussiennes G, chacune est divisée en deux gaussiennes dont les moyennes sont perturbées par un vecteur écart-type σ_k qui est déduit de la diagonale de la matrice de covariance Σ_k. L'augmentation par perturbation peut être calculée par la formule suivante :

$$\mathcal{N}(w_k, \mu_k, \Sigma_k) = \begin{cases} \mathcal{N}(\frac{w_k}{2}, \mu_k - 0.2\sigma_k, \Sigma_k) \\ \\ \mathcal{N}(\frac{w_k}{2}, \mu_k + 0.2\sigma_k, \Sigma_k) \end{cases} \tag{2.13}$$

2.5.3 Décodage de la parole

Pour le décodage de la parole avec HTK, il faut disposer d'un réseau de phonèmes, d'une grammaire et de l'ensemble des modèles HMMs déjà appris. Le réseau de phonèmes correspond a un ensemble de nœuds ou d'états connectés entre eux par un arc. Ce réseau représente la structure de recherche à partir duquel sera réellement effectué le décodage. Le module HVite de décodage, utilise l'algorithme du passage de jeton (voir algorithme 2.1) en anglais token passing proposé par [YOUNG et collab., 1989] qui est une variante de l'algorithme de Viterbi (voir la section 1.8 et la figure 1.4) compatible avec les contraintes de la reconnaissance de phonèmes connectés.

Algorithme 2.1 : *Passage de jeton (Viterbi)*

1. Initialisation :
À l'instant $t=0$, tous les états initiaux reçoivent un jeton de valeur nulle.
Les autres reçoivent un jeton de valeur infinie.

2. Traitement
Pour $t=1$ à T faire :
Pour tous les états i faire :
Passer une copie du jeton de l'état i vers tous les états connectés j, en incrémentant sa valeur de $b_j(t)+a_{ij}$. ($b_j(t)$ correspond à la probabilité d'émission de la trame t dans l'état j et a_{ij} correspond à la probabilité de transition de l'état i vers l'état j).
Fin Pour
Pour tous les états i faire :
Trouver le jeton de plus petite valeur dans l'état i, éliminer les autres.
Fin Pour
Fin Pour

3. Condition d'arrêt :
Examiner tous les états finaux, le jeton avec la plus petite valeur correspond au meilleur score d'alignement.

Dans l'algorithme, les contraintes linguistiques interviennent entre deux phonèmes. Nous avons utilisé un modèle de langage bigramme, estimé sur les étiquettes des phrases d'apprentissage par l'outil HLStats. La chaîne de phonèmes reconnus par ce décodage est comparé avec la chaîne de phonèmes de référence (noyau de test) en utilisant l'algorithme de programmation dynamique DTW réalisé par l'outil HResults. Ce traitement permet de compter les phonèmes reconnus, omis, substitués ou insérés, afin de calculer le taux de reconnaissance phonétique (Accuracy).

2.5.4 Expériences et résultats

Dans le but d'évaluer notre système de reconnaissance monophone, nous avons testé l'apport des coefficients différentiels du premier puis du deuxième et ensuite du troisième ordre par rapport au cas initial des 13 coefficients statique. Nous travaillons avec des vecteurs Mel cepstraux de dimension d=13 (12 MFCC; E), d=26 (12 MFCC; E; 12 ΔMFCC; ΔE), d=39 (12 MFCC; E; 12 ΔMFCC; ΔE; 12 $\Delta\Delta$MFCC; $\Delta\Delta$E), d=52 (12 MFCC; E; 12 ΔMFCC; ΔE; 12 $\Delta\Delta$MFCC; $\Delta\Delta$E; 12 $\Delta\Delta\Delta$MFCC; $\Delta\Delta\Delta$E) . Les coefficients différentiels sont calculés à partir d'une fenêtre d'analyse de 5 trames en utilisant la formule (1.9). Les 48 modèles HMM monophones ont la même topologie (3 états émetteurs), les probabilités d'émission de tous les états sont représentées par une combinaison linéaire de G gaussiennes (GMM) à matrice de covariance diagonale. Le nombre de gaussiennes G est augmenté progressivement (2,4,8,16 jusqu'à 32 gaussiennes par état) grâce à l'outil HHEd. Les modèles sont enfin ré-estimés avec l'outil HERest. Ces modèles HMM monophones et le modèle de langage bigramme sont appris sur les 8 phrases "si" et "sx" des 462 locuteurs d'apprentissage de la base TIMIT, au total 3696 phrases contenant 140225 segments acoustiques. Le décodage est effectué en utilisant le regroupement en 39 classes phonétique de Kai Fu Lee. Les taux de reconnaissance de phonèmes sont représentés dans le tableau 2.6 pour les quatre expériences décrites ci-dessus en fonction du nombre de gaussiennes utilisées dans chaque état.

Nous obtenons les meilleurs résultats en utilisant d=39 coefficients et G=32 gaussiennes par état (voir la figure 2.1). Au delà de 2 dérivées (Δ et $\Delta\Delta$), les performances du système deviennent moins bonnes et le temps de calcul devient conséquent (puisque le nombre de paramètres augmente). L'apport des coefficients différentiels respectivement du premier et second ordre est majeur, environ 9.63% et 13.14%. Le système de reconnaissance

est plus performant avec l'utilisation de 39 coefficients. Cette dimensionnalité représente le nombre référence dans notre système de base.

Nombre de gaussiennes	**Accuracy (%)**	**Correct (%)**
d=13 (12 MFCC ; E)		
1	49.55	51.81
2	51.89	53.92
4	53.64	55.61
8	54.54	56.47
16	55.48	57.44
32	57.05	58.77
d=26 (12 MFCC ; E ; 12 ΔMFCC ; ΔE)		
1	55.70	59.14
2	59.78	62.79
4	62.87	65.49
8	64.48	67.11
16	65.53	68.14
32	66.68	69.06
d=39 (12 MFCC ; E ; 12 ΔMFCC ; ΔE ; 12 $\Delta\Delta$MFCC ; $\Delta\Delta$E)		
1	57.99	62.99
2	61.52	66.44
4	64.60	68.62
8	66.86	70.38
16	68.01	71.37
32	70.19	73.44
d=52 (12 MFCC ; E ; 12 ΔMFCC ; ΔE ; 12 $\Delta\Delta$MFCC ; $\Delta\Delta$E ; 12 $\Delta\Delta\Delta$MFCC ; $\Delta\Delta\Delta$E)		
1	56.40	62.79
2	59.70	66.14
4	63.15	68.75
8	65.59	70.77
16	67.78	72.18
32	68.97	73.18

TABLEAU 2.6: *L'apport des coefficients différentiels sur les taux de reconnaissance de la partie noyau de test (core test) de la base de données TIMIT .*

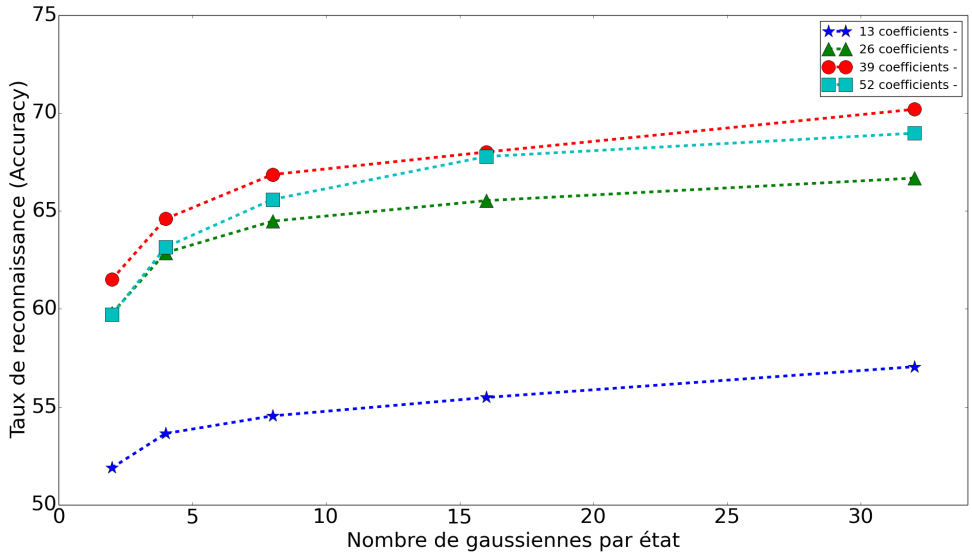

FIGURE 2.1: *L'apport des coeffcients différentiels sur le taux de reconnaissance phonétique (Accuracy) en fonction du nombre de gaussiennes utilisées dans chaque état*

2.6 L'apport du modèle de langage bigramme

L'introduction du modèle de langage permet de déterminer quelles sont les séquences de phonèmes les plus probables au sein de la langue modélisée. C'est une manière d'introduire des informations de nature linguistique. Nous utilisons un modèle de langage bigramme estimé par les outils HLStats et HBuild à l'aide de la transcription phonétique des phrases de l'ensemble de la partie apprentissage de la base de données TIMIT. Ce modèle bigramme est utilisé dans le processus de décodage par HVite pour augmenter la performance de notre système de reconnaissance monophone. L'apport du langage bigramme au décodage est évalué par comparaison avec une expérience de décodage sans bigramme. Le tableau 2.7 donne les taux de reconnaissance avec et sans modèle bigramme obtenus par notre système de reconnaissance monophone de référence. Ce système de référence est appris sur des vecteurs MFCC de 39 coefficients (Δ et $\Delta\Delta$) en utilisant 32 gaussiennes par état. Le gain du taux de reconnaissance apporté par le modèle de langage bigramme est important, de l'ordre de 8%.

Bigramme	Accuracy (%)	Correct (%)
Non	61.87	73.53
Oui	70.19	73.44

TABLEAU 2.7: *L'apport du modèle de langage bigramme sur les taux de reconnaissance de la partie noyau de test (core test) de la base de données TIMIT .*

2.6.1 Facteur d'échelle du modèle de langage

Le facteur d'échelle, est un coefficient introduit dans le processus de décodage au travers du modèle de langage utilisé. Ce facteur est appliqué dans HTK par l'option 's' du module de décodage HVite au niveau des probabilités de transition entre les modèles phonétiques. Une valeur élevé, diminue le nombre d'insertions en pénalisant les transitions entre phonèmes peu fréquentes. Tandis qu'une valeur basse diminue les omissions (phonèmes supprimés). Des expériences montrent l'influence de ce facteur sur la précision du décodage [LJOLJE, 1994; YOUNG et WOODLAND, 1994]. La meilleure valeur de ce facteur dépend fortement des conditions expérimentales. Dans nos expériences, nous avons examiné l'influence de ce facteur par des valeurs comprises entre 1 et 10 sur notre système de reconnaissance monophone HTK à l'aide d'un modèle de langage bigramme. Le taux de reconnaissance de phonème (Accuracy) atteint un maximum pour un facteur d'échelle s=4 (voir le tableau 2.8).

Facteur d'échelle	Phonèmes substitués (%)	Phonèmes supprimés (%)	Phonèmes insérés (%)	Accuracy (%)	Correct (%)
1	20.55	5.68	7.62	66.14	73.76
2	19.42	6.48	5.11	69.00	74.10
3	19.04	7.31	3.92	69.72	73.64
4	18.41	8.15	3.26	70.19	73.44
5	18.13	9.30	2.72	69.85	72.57
6	17.98	10.15	2.38	69.49	71.88
7	17.76	11.10	2.09	69.04	71.13
8	17.81	11.82	1.81	68.55	70.37
9	17.77	12.51	1.59	68.14	69.73
10	17.79	13.33	1.51	67.36	68.87

TABLEAU 2.8: *L'apport du facteur d'échelle du modèle de langage bigramme (résultats obtenus sur le noyau de test (core test) de la base de données TIMIT).*

2.7 Système de reconnaissance triphone

Le même phonème est prononcé différemment selon son contexte. La variabilité du signal de la parole n'est pas parfaitement représentée par les modèles HMM indépendants du contexte (monophones). Afin de prendre en considération les effets liés aux phénomènes de coarticulation plusieurs modèles contextuels ont été proposés. Les auteurs dans [LEE et HON, 1989; LEE et collab., 1990; LJOLJE, 1994] ont prouvé que les taux de reconnaissance de la parole peuvent être nettement améliorés grâce à ces modèles. Il est préférable de travailler avec les modèles triphones tenant compte des contextes phonétiques gauche et droit. Par exemple, la notation HTK du triphone [a]-[l]+[o] signifie que le phonème courant [l] est précédé du phonème [a] et suivi de [o]. Pour un ensemble initial de 48 phonèmes, il existe 48^3 = 110592 triphones possibles. La taille de la base de données phonétiques d'apprentissage peut alors devenir insuffisante pour apprendre correctement chacun des modèles. De plus, un certain nombre de triphones peut ne pas être rencontré dans cette base de données. Pour contourner cette difficulté, il faut d'abord supprimer les triphones non représentés dans la base de données. Deuxièmement, il faut réduire le nombre de modèles ou diminuer le nombre de paramètres du système RAP. Pour cette raison, nous appliquons une approche basée sur le partage de données d'apprentissage entre les états des HMMs triphones (en anglais state-tying). Cette méthode proposée dans [YOUNG et collab., 1994; YOUNG et WOODLAND, 1994], consiste à associer le même GMM aux états qui sont acoustiquement proches. Le partage des états peut se faire soit de manière ascendante, soit de manière descendante.

2.7.1 Partage d'états par approche ascendante

L'approche ascendante consiste à regrouper les contextes droits entre eux et les contextes gauches entre eux. Cela signifie que le premier état d'un modèle triphone ne peut être regroupé qu'avec le premier état d'un autre triphone (voir la figure 2.2). Dans ce processus, les modèles HMM triphones initiaux doivent avoir une seule gaussienne par état. La distance $d(i,j)$ entre deux états i et j (ou groupes d'états) est calculée par l'équation suivante :

$$d(i,j) = \sqrt{\frac{1}{d}\sum_{k=1}^{d}\frac{(\mu_{ik}-\mu_{jk})}{\sigma_{ik}\sigma_{jk}}} \tag{2.14}$$

Où d est la dimension des vecteurs acoustiques, μ_{ik} et σ_{ik} sont les $k^{ièmes}$ coefficients

de la moyenne et de la variance de la gaussienne de l'état i. Les deux états qui minimisent cette distance sont réunis dans un seul groupe (cluster). L'algorithme itère sur toutes les paires d'états jusqu'à ce que toutes les distance soient supérieures à un seuil donné. Ensuite tous les groupes d'états ainsi formés sont examinés de façon à vérifier que le nombre d'échantillons dans la partie apprentissage soit suffisant.

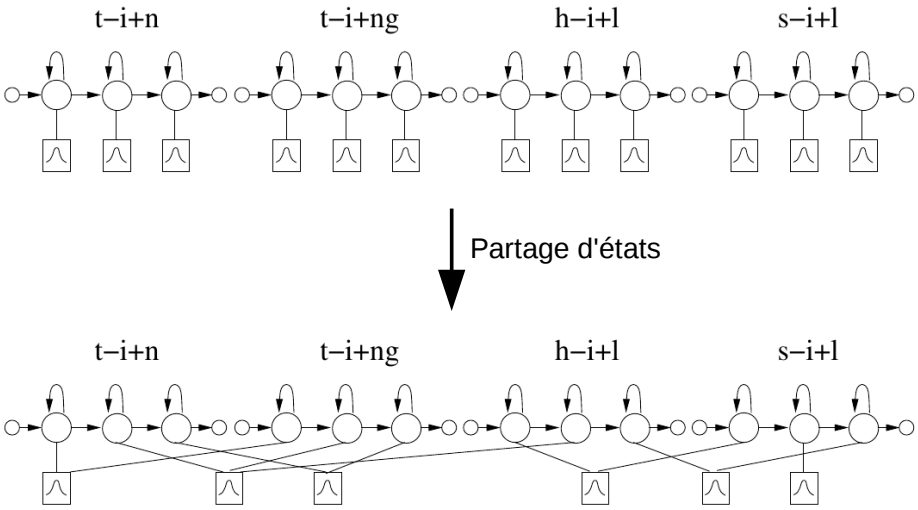

FIGURE 2.2: *Modèles HMM triphones à états partagés.*

2.7.2 Partage d'états par approche descendante

L'approche ascendante ne permet pas de construire un triphone qui n'a jamais été vu lors de l'apprentissage. C'est pour contourner cet inconvénient que l'approche descendante (arbre de décision) a été proposé par [YOUNG et collab., 1994]. Cette approche s'appuie sur des connaissances linguistiques en exploitant un arbre de décision spécifique à chaque état. Une question linguistique binaire est posée à chaque nœud de l'arbre qui porte sur le contexte phonétique gauche ou droit du phonème pris en compte. Par exemple, dans la figure 2.3 la question "est-ce que le phonème suivant (contexte droit) du phonème courant [aa] est une consonne ?" est associée au nœud racine de l'arbre de décision. Une arbre est créé pour chaque état de chaque phonème pour regrouper tout les états similaires des triphones. Deux états fournissant la même réponse sur toutes les questions de l'arbre, partageront les même paramètres. Ces questions linguistiques sont

choisies de façon à maximiser la vraisemblance des modèles avec les données d'apprentissage.

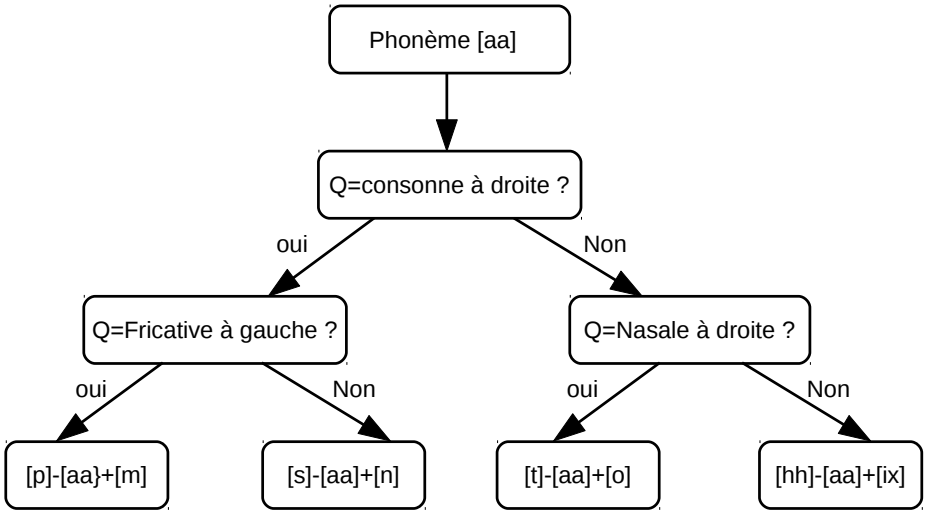

FIGURE 2.3: *Exemple d'arbre de décision utilisé pour partager les états des modèles HMM triphones.*

La vraisemblance totale se présente sous la forme suivante :

$$\mathrm{L(E)} = -\frac{1}{2}(\log[(2\pi)^d|\Sigma(\mathrm{E})|] + d)\sum_{e\in \mathrm{E}}\sum_{t\in \mathrm{T}}\gamma_e(x_t) \tag{2.15}$$

Où E correspond à l'ensemble des états HMM, T le nombre de trames (vecteurs acoustique) et $\gamma_e(x_t)$ est la probabilité a posteriori que le vecteur x_t soit généré par l'état e calculé en utilisant l'algorithme de Baum-Welch, d est la dimension des vecteurs.

En pratique l'algorithme de partage d'états par arbre décision (voir algorithme 2.2) réduit le nombre d'états sans aucune dégradation en performance.

Algorithme 2.2 : *partage d'états par arbre de décision*

1. Regrouper tous les contextes phonétiques en une seule classe.

2. Trouver le nœud N et la question Q qui maximisent l'accroissement de la vraisemblance obtenu en partitionnant les états en deux sous ensemble $E_o(q)$ et $E_n(q)$.
   tel que : $\Delta L_Q = L(E_o(q)) + L(E_n(q)) - L(E)$ (avec o :oui, n :non)

3. Si ΔL_Q dépasse un seuil donné, alors on découpe N selon Q, et en recommence à l'étape 2, sinon on continue.

4. Trouver les nœud N_1 et N_2 qui minimisent la diminution de la vraisemblance lorsque les paramètres (moyenne et matrice de covariance) sont fusionnées.

5. Si cette diminution est inférieur à un seuil donné, alors fusionner N_1 et N_2 et itérer à partir de l'étape 4, sinon continuer.

2.7.3 Expérience et résultats

La première étape pour construire des modèles HMM triphones est d'utiliser un simple clonage des modèles indépendant du contexte (monophones) déjà appris. Les vecteurs moyens et les matrices de covariances, ainsi que les probabilités de transition seront identiques pour tous les triphones associés au monophone approprié. L'outil HLEd d'HTK, permet de générer la liste de tous les triphones pour lesquels il existe au moins un exemple dans la partie apprentissage de la base de données. Les 48 modèles monophones que nous avons utilisés pour créer les triphones sont appris avec 1 gaussienne/état avec des vecteurs MFCC de dimension d = 39 représentant la configuration de référence. L'ensemble des modèles triphones créés doit être ré-estimé en utilisant l'outil HERest. Pour faire cela, la transcription des données d'apprentissage doit être convertie en étiquettes dépendant du contexte (voir la figure 2.4).

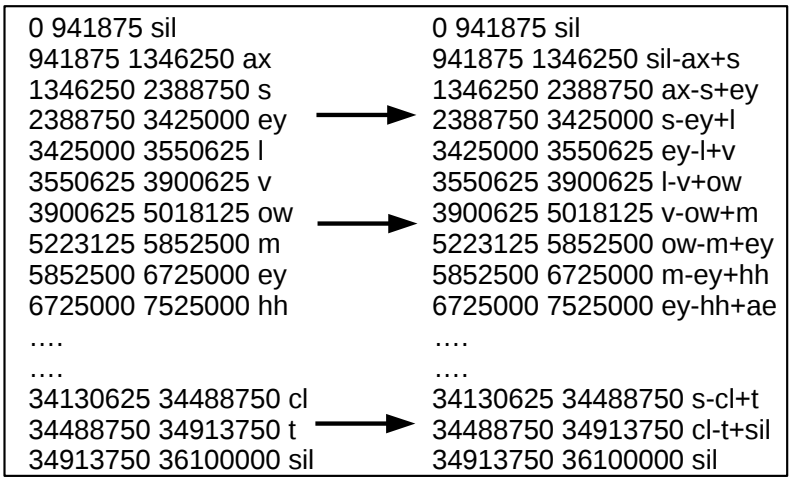

FIGURE 2.4: *conversion de la transcription monophones en transcription triphones du fichier dr1/fcjf0/si648.lab*

Le nombre de modèles HMM passe alors de 48 monophones à plusieurs milliers de triphones. Il est impensable de disposer de données suffisantes pour faire un apprentissage correct de la totalité de ces modèles triphones. En effet, certains n'apparaissent que quelque fois dans la base d'apprentissage. Pour contourner cette difficulté, nous avons choisi d'utiliser la méthode de partage d'états par arbre de décision décrite dans la section précédente. Les arbres de décision sont donc construits pour chaque classe phonétique en utilisant une procédure d'optimisation séquentiel de haut en bas. Initialement tous les modèles triphones appartenant à la même classe phonétique sont placés dans un seul groupe à la racine de l'arbre. Une série de questions linguistiques binaires (QS) générée par le script 'mkclscript' d'HTK est exécuté pour partitionner les états qui maximisent la vraisemblance. Le processus de partitionnement est répété jusqu'à ce que l'augmentation de cette vraisemblance tombe en dessous d'un seuil (TB) spécifié. En phase finale, toutes les paires d'états pour lesquelles la diminution de la vraisemblance est inférieure au seuil utilisé pour arrêter le partitionnement sont ensuite fusionnées. Un autre seuil (RO) des valeurs anormales est utilisé pour supprimer les triphones qui n'ont pas suffisamment de données pour être ré-estimés. Ce seuil est lié aux statistiques d'occupation minimal des groupe d'états. Nous avons fait varié les valeurs des seuil RO de 100 à 190 par pas de 30 et le seuil TB de 400 à 800 par pas de 200. Il faut noter que les valeurs des seuil RO et TB affectent le degré de regroupement (liaison) des états et donc le nombre final des états et des modèles triphones. Les valeurs doivent être modifiées suivant la quantité de données

d'apprentissage disponible.

Certains modèles triphones peuvent partager exactement les 3 mêmes états émetteurs et les matrices de covariances et de transitions et sont donc identiques. Dans ce cas les deux modèles triphones identiques sont regroupés ensemble par confusion dans un même modèle HMM. Après avoir ré-estimé les modèles triphones créés, le nombre de gaussiennes est ensuite augmenté itérativement de 2,4,8 jusqu'à 16 gaussiennes par état en utilisant l'outil HHEd (il n'y a pas assez de données pour faire un apprentissage des triphones avec 32 gaussiens par état). A chaque itération les modèles triphones sont ré-estimés en utilisant le nombre de gaussiennes attribué. Le tableau 2.9, illustre l'effet de faire varier les deux seuil RO et TB sur le nombre de modèles triphones créés et le nombre d'états final ainsi que sur les taux de reconnaissance en utilisant 16 gaussiennes dans chaque état.

Seuils	Nombre de triphones	Nombre d'états	Accuracy (%)	Correct (%)
RO=100, TB=400	5870	1490	72.34	76.59
RO=100, TB=600	3745	1045	71.68	75.95
RO=100, TB=800	2470	823	71.55	75.55
RO=130, TB=400	5628	1457	72.27	76.60
RO=130, TB=600	3715	1040	72.57	76.56
RO=130, TB=800	2467	821	71.93	75.80
RO=160, TB=400	5561	1429	72.27	76.48
RO=160, TB=600	3686	1026	**72.64**	**76.59**
RO=160, TB=800	2459	819	71.81	75.68
RO=190, TB=400	5361	1401	71.91	76.27
RO=190, TB=600	3470	1013	72.58	76.47
RO=190, TB=800	2400	815	72.18	75.94

TABLEAU 2.9: *Le nombre de modèles triphones et groupes d'états pour les différentes valeurs des seuils RO et TB, ainsi que les taux de reconnaissance obtenus sur la partie core test de la base de données TIMIT.*

Le meilleur taux de reconnaissance phonétique (Accuracy) est atteint en utilisant 3686 triphones et 1026 états partagés avec 16 gaussiennes par état. Cette configuration est générée par les seuils RO=160 et TB=600 (voir tableau 2.9). Il est à noter que le facteur d'échelle du modèle de langage bigramme est modifié à 8 au lieu de 4 pour les modèles

monophones (les performance du système triphone diminue avec des valeurs inférieures ou supérieures à 8).

2.8 Réduction de la dimensionnalité et discrimination des vecteurs acoustiques

Il est évident que les performances d'un système RAP s'améliorent par l'utilisation des coefficients différentiels du premier et second ordre (Δ et $\Delta\Delta$). Cependant, ces coefficients entraînent un triplement de la taille des vecteurs acoustiques et manquent de discrimination au niveau de ces paramètres. Il est donc préférable de ne conserver que les coefficients discriminants et réduire la redondance de l'information présente. Divers techniques ont été proposées pour effectuer cette tâche, comme l'Analyse en Composantes Principales (ACP) pour décorréler les coefficients [TOKUHIRA et ARIKI, 1999], l'Analyse Linéaire Discriminante (en anglais : LDA pour Linear Discriminant Analysis) et son extension Heteroscedastic LDA (HLDA).

Nous allons décrire dans la suite ces deux techniques permettant de transformer les vecteurs acoustiques dans un espace de dimension restreint possédant de bonnes propriétés discriminantes. Nous avons implémenté la méthode HLDA pour améliorer la performance de notre système de reconnaissance (triphones).

2.8.1 Analyse Discriminante Linéaire (ADL)

L'analyse discriminante linéaire [HAEB-UMBACH et NEY, 1998], est une méthode de réduction de la dimension qui consiste à projeter les vecteurs acoustiques $\mathrm{X}_\mathrm{N}^d = [x_1^d, x_2^d, \ldots, x_\mathrm{N}^d]$ de l'espace R^d dans un sous-espace R^p plus petit ($p \leq d$), de manière à maximiser la discrimination entre les classes. Cette projection est accomplie mathématiquement par la transformation linéaire suivante :

$$\mathrm{Y}^p = \Theta_p^d \mathrm{X}^d \tag{2.16}$$

Où Θ représente la matrice de transformation de dimension (pxd) et Y^p les vecteurs transformés dans l'espace discriminant de p coefficients. La procédure d'analyse discriminante consiste à chercher la matrice de transformation optimale $\tilde{\Theta}$ en maximisant la

variance inter-classes et en minimisant la variance intra-classes par le critère suivant :

$$\tilde{\Theta} = \arg\max_{\Theta_p} \left(\frac{\Theta_p^d \mathrm{S_B} \Theta_p}{\Theta_p^d \mathrm{S}_w \Theta_p} \right) \tag{2.17}$$

- $\mathrm{S_B}$ correspond à la matrice de covariance inter-classes :

$$\mathrm{S_B} = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} (x_i - \mu)(x_i - \mu)' \tag{2.18}$$

avec $\mu = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} x_i$ le vecteur moyen global et N le nombre total de vecteurs

- $\mathrm{S_W}$ correspond à la matrice de covariance intra-classes :

$$\mathrm{S}_w = \sum_{j=1}^{c} \left(\frac{1}{\mathrm{N}_j} \sum_{i=1}^{\mathrm{N}_j} (x_i - \mu_j)(x_i - \mu_j)' \right) \tag{2.19}$$

avec $\mu_j = \frac{1}{\mathrm{N}_j} \sum_{i=1}^{\mathrm{N}_j} x_i$ est le vecteur moyen de la classe j, N_j le nombre de vecteurs de la classe j et c le nombre total de classes.

La solution de l'équation 2.17 peut être trouvée par le calcul des vecteurs propres de la matrice $\mathrm{S_W^{-1}S_B}$. L'ADL ou LDA (en anglais) est couramment employée dans le domaine de la reconnaissance automatique de la parole [HAEB-UMBACH et NEY, 1998; SIOHAN, 1995] afin d'améliorer la discrimination des vecteurs acoustiques.

2.8.2 Héteroscedastic LDA (HLDA)

Héteroscedastique LDA (HLDA) est une variante de la technique LDA. LDA suppose que la moyenne est le facteur discriminant et non la variance, car les distributions des classes sont gaussiennes avec des vecteurs moyens différents et matrices de covariance communes (Homoscédasticité). En raison de cet inconvénient, LDA peut fournir des performances insatisfaisantes lorsque les distributions de classe sont hétéroscédastiques (variances ou covariances inégales). C'est pour remédier à cette limitation que la transformation HLDA [KUMAR et ANDREOU, 1998] a été proposée. Le principe de la transformation HLDA est un peu différent par rapport à la technique LDA. La matrice de transformation Θ est étendue à dxd dimensions.

$$\mathrm{Y} = \Theta.\mathrm{X} = \begin{bmatrix} \Theta_p \mathrm{X}^d \\ \Theta_{d-p} \mathrm{X}^d \end{bmatrix} = \begin{bmatrix} \mathrm{Y}^p \\ \mathrm{Y}^{d-p} \end{bmatrix} \tag{2.20}$$

Où Θ_p représente les p première lignes de la matrice de transformation Θ et Θ_{d-p} les $d-p$ lignes restantes. Chaque classe j est modélisée par une distribution normale des X_n vecteurs d'apprentissage (d'entrée).

$$p(x_i) = \frac{|\Theta|}{\sqrt{(2\pi)^d |\Sigma_j|}} \exp(-\frac{1}{2}(\Theta x_i - \mu_j)' \Sigma_j^{-1} (\Theta x_i - \mu_j)) \tag{2.21}$$

Où μ_j, Σ_j représentent (respectivement) le vecteur moyen et la matrice de covariance de la classe j. Le but est de déterminer la solution optimale qui respecte un critère de maximisation de la fonction de probabilité log-vraisemblance des données en terme de Θ.

$$\tilde{\Theta} = \arg\max_{\Theta} \sum_{\forall i} \log(p(x_i) \tag{2.22}$$

L'algorithme itératif efficace proposé dans [BURGET, 2004; GALES, 1999], basé sur une version généralisée de l'algorithme EM est utilisé dans nos expériences pour simplifier l'estimation de la matrice $\tilde{\Theta}$. Une fois la matrice optimale de transformation $\tilde{\Theta}$ obtenue, les p première lignes de cette dernière sont utilisées pour calculer les vecteurs discriminants Y^p par la projection 2.16.

Nous avons effectué 2 expériences sur notre système triphone, afin d'évaluer l'apport de la transformation HLDA des vecteurs acoustiques MFCC sur le taux de décodage de la parole. Dans la première expérience nous avons utilisé des vecteurs MFCC de dimension d=39 (12 MFCC; E; 12 ΔMFCC; ΔE; 12 $\Delta\Delta$MFCC; $\Delta\Delta$E), qui représentent le cas de référence (la meilleure configuration de notre système). Ces vecteurs de 39 coefficients ne subissent pas de réduction de dimension mais ils sont transformés dans un espace plus discriminant (39 → 39). Dans la deuxième expérience, la matrice de transformation HLDA de dimension (39x52) est calculée sur des vecteurs MFCC de dimension d=52 (12 MFCC; E; 12 ΔMFCC; ΔE; 12 $\Delta\Delta$MFCC; $\Delta\Delta$E; 12 $\Delta\Delta\Delta$MFCC; $\Delta\Delta\Delta$E) incluant la troisième dérivée ($\Delta\Delta\Delta$). Ensuite l'espace de 52 coefficients est réduit à 39 par l'application de la matrice de transformation HLDA sur les vecteurs MFCC (52 → 39).

les taux de reconnaissance obtenus sont représentés dans le tableau 2.10 pour les 2 expériences décrites ci-dessus sur la partie noyau de la partie test de la base de données TIMIT.

3686 triphones (RO=160, TB=600) 1026 groupe d'états avec 16 Gaussiennes par état + Bigramme	**Accuracy (%)**	**Correct (%)**
Expérience 1 : HLDA coefficients (39 → 39)	71.24	77.98
Expérience 2 : HLDA coefficients (52 → 39)	**74.91**	**78.23**
Sans HLDA : 39 coefficients	72.64	76.59

TABLEAU 2.10: *L'apport des coefficients différentiels et de la transformation HLDA sur le taux de reconnaissance phonétique (Accuracy) obtenu sur la partie core test de la base de données TIMIT.*

Nous remarquons, que le taux de reconnaissance phonétique (Accuracy) est amélioré de 2.27% par rapport au cas normal (sans transformation HLDA).

2.9 Conclusion

Nous avons construit trois systèmes (SPIRIT, monophone, triphone) de RAP continue indépendants du locuteur. Chacun d'eux comporte ses propres caractéristiques et méthodes de modélisation, d'apprentissage et de test. L'évolution des performances se déroule d'une façon progressive lors du passage d'un système à l'autre. Les meilleurs résultats sont obtenus grâce au système de reconnaissance triphone utilisant des modèles phonétiques dépendants du contexte. Ce système prend en considération l'expertise actuelle en matière de reconnaissance de la parole, et présente une qualité de décodage tout a fait satisfaisante par rapport à d'autre systèmes à base d'HMM. En plus, nous avons transformé les vecteurs acoustiques MFCC à l'aide de la méthode HLDA pour maximiser l'information discriminante entre les classes phonétiques. Le taux de reconnaissance de phonème (Accuracy) est de 74.91% obtenu sur la partie noyau de la partie test de la base de données TIMIT.

Chapitre 3

Reconnaissance automatique de la parole alaryngée

« *La vie est un mystère qu'il faut vivre, et non un problème à résoudre.* »

Gandhi

3.1 Introduction

La reconnaissance et l'évaluation de la parole alaryngée (pathologique), est l'un des sujets sensibles au centre de nombreuses études dans des domaines multi-disciplinaires [DIBAZAR et collab., 2006; PRAVENA et collab., 2012]. La parole pathologique, désigne la parole produite par des locuteurs atteints de dysfonctionnement (altération du son laryngé) de la voix et de la parole. Le dysfonctionnement vocal peut être évalué, soit par des jugements de perception ou par une analyse objective.

L'analyse par des jugements de perception est la méthode incontournable, la plus utilisée en pratique clinique. Elle consiste à caractériser la qualité vocale par une simple écoute attentive. Toutefois, cette technique souffre de plusieurs inconvénients. Tout d'abord, le jugement perceptuel doit être effectué par un jury d'experts en vue d'accroître sa fiabilité. Deuxièmement, cette analyse perceptuelle est très coûteuse en temps et en ressources humaines et ne peut être planifiée régulièrement.

De nos jours, l'analyse objective [WUYTS et collab., 2000; YU et collab., 2001] est de plus en plus utilisée. Elle se base sur l'analyse des mesures acoustiques, aérodynamiques et physiologiques. Ces mesures peuvent être directement extraites du signal de la parole à l'aide d'un système informatique. Cette approche objective offre des résultats acceptables mais encore insuffisants pour la reconnaissance automatique et l'évaluation de la parole œsophagienne. Face à ces faiblesses, nous avons proposé une méthode instrumentale à la fois simple et rapide pour décoder et évaluer la parole œsophagienne en appliquant un système RAP continue (phonèmes connectés) sur notre propre base de données de la parole œsophagienne FPSD (French Pathological Speech Database).

Dans ce qui suit, quelques notions sur la parole pathologique seront présentées. Ensuite nous décrirons notre corpus de la parole œsophagienne FPSD ainsi que notre méthode proposée pour l'évaluation et le décodage de la parole œsophagienne.

3.2 Parole pathologique

La parole pathologique provient de certains troubles de la voix, qui se traduisent par une modification au niveau des paramètres acoustiques (altération objective) ou/et sonores (altération subjective) de la parole. Ce dysfonctionnement de la voix peut être momentané ou durable.

En général, il existe trois grandes catégories de pathologies :

A) **Les pathologies d'origines fonctionnelles :** mauvaise utilisation des organes de la phonation (conduit vocal), la cause est souvent liée à l'âge du patient (locuteur). On retrouve parfois une altération de la voix de cause psychologique comme par exemple, une dépression.

B) **Les pathologies d'origines organiques :** laryngite aiguë, présence de lésion sur les cordes vocales, kystes, etc. Les principales causes de ces pathologies sont le forçage de la voix et les infections virales ou bactériennes du larynx.

C) **Les pathologies d'origines cancereuses :** l'ablation partielle ou totale du larynx est un acte chirurgical motivé par un cancer. La consommation d'alcool et l'usage du tabac en sont les principales causes.

Dans cette thèse, nous étudierons les dysfonctionnements de la voix dus aux pathologies d'origines cancéreuses.

3.2.1 Le cancer du larynx

Le larynx (figure 3.1) comporte plusieurs organes. Il se trouve au carrefour des voies aériennes et digestives, entre le pharynx et le trachée, et en avant de l'œsophage. Les cordes vocales sont deux lèvres symétriques (structure fibreuse) placées au travers du larynx. Le passage de l'air expiratoire provenant des poumons lors de la phonation, met en vibration la muqueuse des cordes vocales en adduction, ce qui permet de produire un son vocal de qualité à l'aide de l'amplification du conduit vocal.

Le cancer du larynx est caractérisé par une tumeur de la forme d'une ulcération anormale d'une des deux cordes vocales. Le traitement consiste alors en une radiothérapie et une chimiothérapie, associée à l'ablation de la corde vocale atteinte (cordectomie). Ce-

1. Illustration extraite de : http://lecerveau.mcgill.ca (sous copyleft)

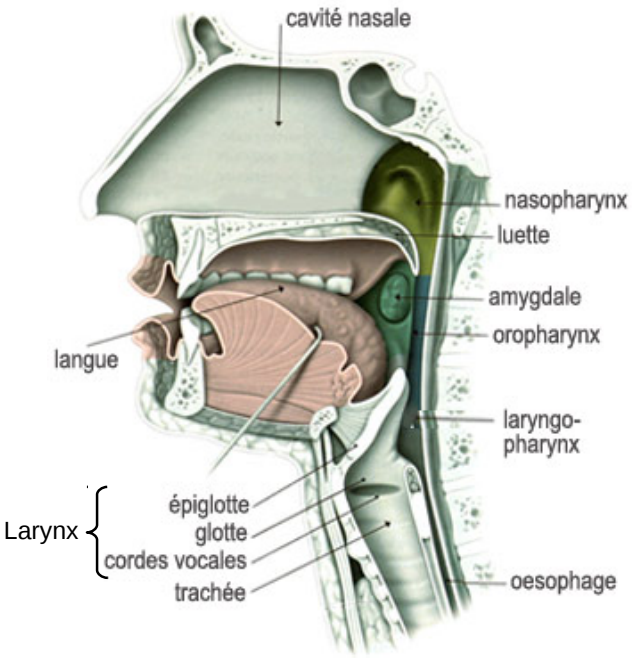

FIGURE 3.1: *Vue shématique des organes de l'appareil vocal*[1]

pendant, dans le cas d'une récidive ou lorsque l'entendu du cancer est trop important et touche la quasi totalité de l'organe, l'ablation complète du larynx (laryngectomie totale) est nécessaire.

Le cancer du larynx est une pathologie tumorale relativement fréquente chez les hommes. D'après les dernières statistiques publiées par l'institut de veille sanitaire [2], il représente en France, environ 25% des atteintes cancéreuses des voies aéro-digestives supérieures et 15% de l'ensemble des cancers diagnostiqués. Au Maroc , d'après le service d'épidémiologie de l'institut National d'oncologie de Rabat entre 1985 et 2007, le cancer du larynx représente 30.8% des cancers du système respiratoire et 9.2% de l'ensemble des cancers enregistrés. La tranche d'âge la plus touchée chez les hommes est celle de 50 à 54 ans, suivie de celle de 55 à 59 ans. Cette affection touche essentiellement les hommes avec 94% contre 6% seulement de femmes. Le tabagisme actif en est la principale cause, aggravé par la consommation conjointe d'alcool et l'inhalation de matières cancérigènes telle que l'amiante.

2. Statistiques disponibles sur le site Internet de l'institut : http://www.invs.sante.fr

3.2.2 Laryngectomie totale

Une laryngectomie totale est une opération chirurgicale consistant en l'ablation complète du larynx afin de traiter un cancer à l'état avancé. Par conséquent, le patient perd ses cordes vocales et ainsi la voix laryngée. En effet, l'air pulmonaire passe exclusivement par le trachéostome (voir figure 3.2) et ne peut donc pas atteindre la cavité buccale. Sans air, la phonation est impossible. Après la chirurgie, certains patients peuvent renoncer à toute tentative de communication orale en raison du bouleversement physique et mental causé par l'acte chirurgical. En effet les changements anatomiques privent temporairement le patient de sa voix. Seule la voix chuchotée permet la communication dans une vie post-opératoire. Pour la rétablir partiellement, plusieurs techniques existent permettant de lui procurer une nouvelle voix de remplacement ou de substitution.

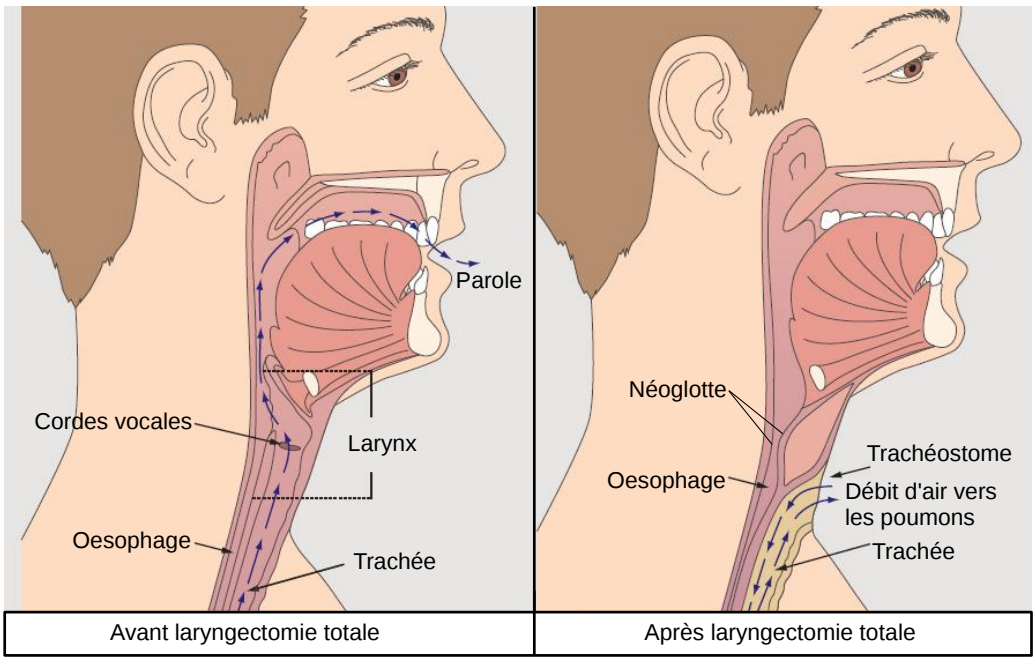

FIGURE 3.2: *Appareil phonatoire d'une personne laryngectomisée (à droite, avant, à gauche, après l'opération).*

3.2.3 Les voix de substitution (réhabilitation vocale)

Laryngectomisé, le patient est contraint d'apprendre une nouvelle voix dite de substitution. En effet la déviation du trajet de l'air pulmonaire due à la suppression de la totalité du larynx empêche ce patient de produire une voix laryngée (normale). L'apprentissage d'une nouvelle voix est permis par les organes bucco-phonatoires ainsi que l'œsophage.

Plusieurs techniques sont proposés au patient après l'opération :

- **La voix œsophagienne :** généralement, la plus utilisée après une laryngectomie totale. À cause du trachéostome (voir la figure 3.2), le patient ne peut plus utiliser l'air provenant des poumons, l'œsophage servira désormais de vibrateur et de réserve d'air : d'où l'appellation de "parole œsophagienne". La voix digestive remplace alors la voix respiratoire dans le rôle de soufflerie. Nous récupérons une analogie avec les trois éléments nécessaires pour la production de la parole : un souffle d'air provenant de l'œsophage, un muscle vibrant appelé "néoglotte" placé dans partie supérieure de l'œsophage et enfin le conduit vocal qui n'a pas subit de changement. Cette voix œsophagienne permet au patient de communiquer d'une manière plus autonome puisqu'elle n'exige aucun outil particulier. Par contre, elle est difficile à maîtriser et longue à acquérir.
- **La voix trachéo-œsophagienne :** cette technique consiste à réaliser une communication entre le trachée et l'œsophage, par la mise en place chirurgicalement, d'une prothèse de silicone (voir la figure 3.3). Cette dernière autorise le passage de l'air provenant des poumons, de la trachée vers l'œsophage et interdit le passage des aliments et des liquides de la cavité buccale vers le trachée. Contrairement à la voix œsophagienne, l'air n'a plus besoin d'être ingéré. En effet, la prothèse phonatoire permet de rediriger l'air pulmonaire depuis la trachée vers l'œsophage pour faire vibrer la néoglotte. La durée possible de phonation est plus longue, et la parole produite est généralement d'une intelligibilité plus satisfaisante. Cependant, la durée de vie de l'implant phonatoire est très limitée, en moyenne de quatre à huit mois ; il devra donc être remplacé deux à trois fois par an. En plus, cette voix trachéo-œsophagienne n'est pas toujours possible et la présence de la prothèse phonatoire peut parfois entraîner des complications (fuite alimentaires autour de la prothèse, déplacement, etc.).
- **La voix Electro-larynx :** est la dernière possibilité proposée à une personne laryngectomisée pour rétablir la communication vocale après l'intervention. Elle est générée par l'intermédiaire d'un appareil portable qui est maintenu contre le menton (voir la figure 3.4). Ce dispositif permet de produire une vibration qui est ensuite modulée par la bouche pour produire une voix synthétisée. Cette technique ne requiert aucun apprentissage, néanmoins la parole produite reste très robotique.

3. Illustration extraite de : http://www.headandneckcancerguide.org/

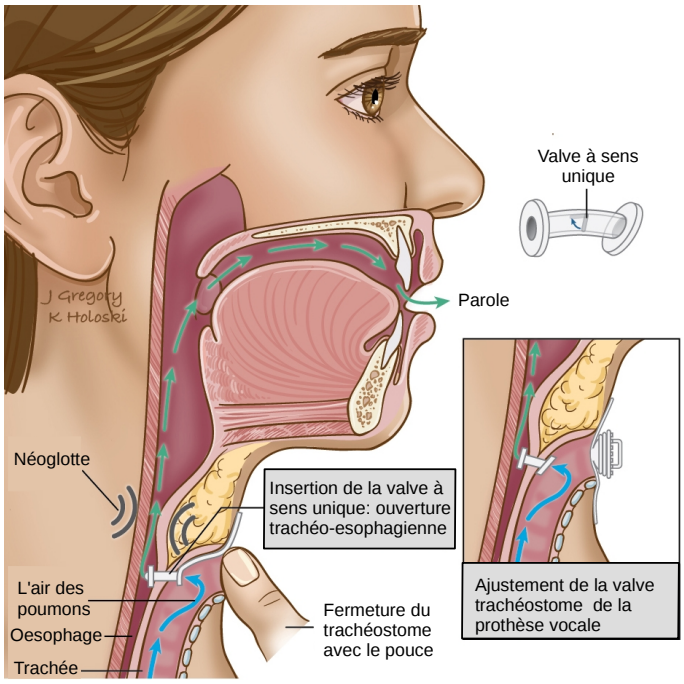

FIGURE 3.3: *Parole trachéo-œsophagienne avec implant phonatoire : en bouchant le trachéostome, l'air passe par l'implant vers l'œsophage et la bouche*[3].

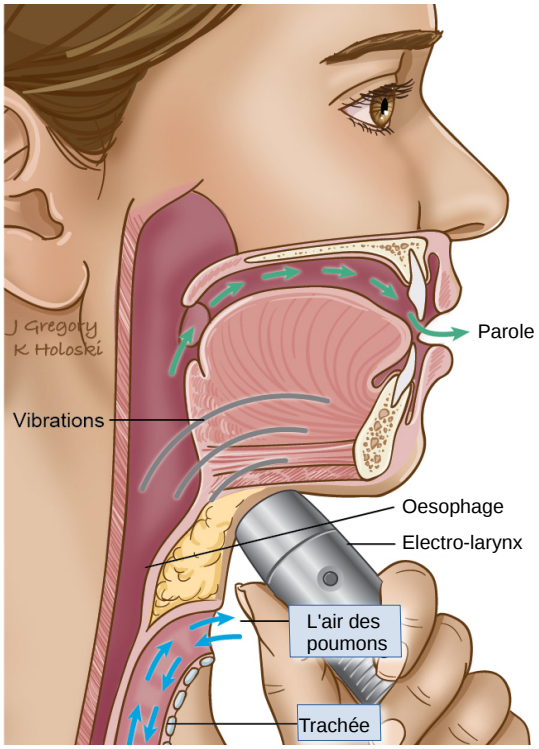

FIGURE 3.4: *Parole electro-larynx à l'aide du dispositif portable*[3].

3.2.4 Caractéristiques acoustiques de la parole pathologique (alaryngée)

Différents travaux de recherche, basés sur le traitement du signal acoustique, ont été effectués pour analyser les caractéristiques acoustiques de la parole alaryngée. Ces études pourront aboutir à des avancées en diagnostics automatiques et à l'établissement de systèmes experts capables de caractériser les anomalies vocales. Les voix de substitution ne peuvent pas être classifiées par des systèmes de classification conçus pour la voix laryngée en raison des propriétés très différentes par rapport à celles de la voix normale :

- **Voisement :** La qualité de la parole alaryngée est influencée par le changement du mécanisme de voisement. Ce changement a des effets sur les différentes caractéristiques acoustiques de la parole. D'abord, la F0 d'une voix de substitution est instable avec une fréquence et un rapport harmoniques/bruit HNR (Harmonics to Noise Ratio) significativement inférieurs à celui de la parole laryngée.
- **Voix Electro-larynx :** cette voix semble très mécanique en raison du signal d'excitation monotone, qui est strictement périodique avec un pitch constant. Un autre sérieux problème est observé dans le son direct rayonné de l'appareil à l'auditeur, est la présence d'un bruit de fond constant [Carol et collab., 1998]. Des études antérieures ont montré que le lissage du contour de la F0 diminue l'intelligibilité des phrases prononcées par des locuteurs sains [Laures et Bunton, 2003; Laures et Weismer, 1999]. Le son robotique de la parole électro-larynx est dû au manque de composantes basses fréquences inférieures à 500 Hz [Qi et Weinberg, 1991].
- **Voix œsophagienne et trachéo-œsophagienne :** le signal d'excitation produit par la néoglotte (vibrateur) est souvent irrégulier, ce qui se manifeste par une voix très rauque. L'enveloppe de la forme d'onde et les composantes spectrales de la parole œsophagienne ne varient pas aussi bien que ceux de la parole laryngée. Par ailleurs, le pitch de la parole œsophagienne est plus faible et moins stable que celui de la parole laryngée. Par conséquent, le processus d'analyse et d'extraction du F0 échoue. L'étude proposée dans [Bellandese et collab., 2001] a dévoilée qu'il existe une différence significative relative à la fréquence fondamentale entre la parole laryngée et alaryngée, mais pas entre la parole œsophagienne et trachéo-œsophagienne. En outre, ces deux voix alaryngées sont faibles en intensité et contiennent un bruit spécifique particulièrement élevé. Toutes ces caractéristiques produisent des sons non

naturels et difficiles à comprendre.

- **Formants :** permettent d'étudier les transformations apportées sur le signal de la parole, lors de sa transition à travers les cavités de l'appareil phonatoire. Les valeurs de ces formants ont subi une légère modification (augmentation) [MELTZNER, 2003; REHAN et collab., 2007]. Ceci peut être justifié par le fait que la configuration du conduit vocal a changé (réduite) en raison du retrait du larynx. Ce changement important a pour conséquence la modification de position des formants.
- **Réserve d'énergie :** Seule la parole électro-larynx offre un niveau d'énergie fixe. La parole trachéo-œsophagienne a une provision d'énergie instable. Tandis que pour la parole œsophagienne, la quantité d'air obtenue par éructation est insuffisante (moins de 80 ml) comparée à celle provenant des poumons dans la parole normale laryngée (environ 5000 ml).

3.3 Création de notre base de données FPSD

Les corpus de la parole pathologique sont relativement moins nombreux par rapport à ceux de la parole laryngée. Souvent les analyses portent sur quelques dizaines de phrases enregistrées par des locuteurs laryngectomisés pour des besoins ponctuels d'une étude. L'enregistrement des signaux et le stockage de données acoustiques sont souvent effectués par du personnel non expérimenté pour certains aspects techniques. A cela s'ajoute la perte fréquente des métadonnées comme par exemple le type de voix pathologique (voix œsophagienne, trachéo-œsophagienne ou electo-larynx, l'âge du locuteur laryngectomisé, le contexte d'enregistrement : analyse, reconnaissance automatique, etc.). C'est pour ces raisons que nous avons choisi de concevoir notre propre base de données française de la parole œsophagienne intitulée FPSD "French Pathological Speech Database".

3.3.1 Configuration de l'enregistrement

Notre corpus acoustique et phonétique FPSD est destiné à la reconnaissance automatique de la parole œsophagienne. Elle contient les enregistrements sonores de 480 phrases différentes prononcées par un seul locuteur mâle âgé de 55 ans qui a subi une laryngectomie totale. Ce locuteur laryngectomisé a acquis la voix œsophagienne après une rééducation vocale grâce à la technique d'éructation contrôlée qui a duré plusieurs mois.

Les 480 phrases prononcées, sont classifiées en cinq catégories :

C1) Phrases avec des mots d'une syllabes.

C2) Phrases avec des mots d'une et deux syllabes.

C3) Phrases avec des mots de trois syllabes.

C4) Phrases d'intonation descendante.

C5) Phrases d'intonation montante.

Les enregistrements sonores ont été effectués par le patient laryngectomsé lui-même. Le signal sonore a été échantillonné à 16 KHz avec 16 bits par échantillon et directement stocké dans des fichiers de type wave sur un ordinateur. L'objectif principal était d'enregistrer une quantité phonétique conséquente afin de faciliter l'implémentation d'un système de reconnaissance automatique de la parole œsophagienne.

3.3.2 Structure du corpus FPSD

Il est nécessaire d'avoir un assez grand corpus d'apprentissage afin de traiter toute la variabilité intra-locuteur. Le plus important est de simplifier le développement d'un système de reconnaissance automatique de la parole œsophagienne. C'est pourquoi, nous avons divisé notre base de données en deux parties : une pour l'apprentissage contenant 425 phrases et l'autre pour le test contenant 55 phrases. La structure des fichiers de notre base de données FPSD est semblable à celle utilisée dans la base TIMIT [GAROFOLO et collab., 1993]. Nous disposons pour chaque phrase, d'un fichier wave (.wav) contenant le signal sonore, d'un fichier texte (.txt) contenant le texte français, d'un fichier (.wrd) contenant la transcription en mots, et d'un fichier (.phn) contenant la segmentation manuelle en phonèmes.

3.3.3 Étiquetage et segmentation manuelle en phonèmes

La segmentation de la parole en phonèmes consiste à délimiter le signal acoustique d'une phrase donnée en séquence de segments. Chaque segment possède ses propres propriétés qui permettent de le différencier des autres. Il est caractérisé par une étiquette de l'alphabet phonétique de la langue modélisée.

Le signal de la parole peut être segmenté en mots et en phonèmes par deux méthodes : soit manuellement par un expert humain, soit d'une façon automatique à l'aide d'une analyse programmée. Qualitativement, la segmentation manuelle est la plus précise. En effet, bien qu'il soit difficile d'évaluer la qualité d'une segmentation phonétique, un consensus a conclu au fait qu'une segmentation manuelle est plus correcte qu'une segmentation automatique. Cependant, cette segmentation manuelle est une tâche très lourde, très longue et difficile à mettre en œuvre même pour la parole laryngée (normale) car les segments constituant le signal de la parole ne sont pas clairement bien délimités. A cela, s'ajoute les diversités de caractéristiques existant entre la parole œsophagienne et la parole laryngée (voir la section 3.2.4). En effet, le phénomène de coarticulation de la parole œsophagienne par la transition d'un phonème à un autre se fait d'une manière bruité avec un chevauchement anormal et étendu. Tous ces inconvénients rendent la tâche de segmentation plus compliquée même pour une oreille humaine (difficulté de perception et de décodage). Il nous fallu environ 4 mois de travail intensif, avec une moyenne de 4 phrases par jour pour pouvoir segmenter manuellement les 480 phrases de notre base de données FPSD.

Certains critères de base ont été utilisés pour perfectionner cette segmentation manuelle :

- **La visualisation du spectrogramme** du signal de la parole facilite la distinction des régions spectralement homogènes en se basant sur les caractéristiques acoustiques propre à chaque son (phonème). Le spectrogramme est une représentation tridimensionnelle, où le temps est représenté sur l'axe des abscisses, la fréquence sur l'axe des ordonnées et le niveau d'amplitude est symbolisé par le niveau de gris. La fréquence, exprimée en Hertz (Hz), est le nombre de répétition d'une période par seconde. Plus elle est élevée plus le son paraîtra "aigu", à l'inverse, il paraîtra "grave". L'amplitude appelée aussi intensité ou volume sonore mesure la pression de l'air en décibels (dB). Un exemple de spectrogramme d'un signal de la parole œsophagienne (FPSD) et de la parole normale (laryngée) pour la même phrase sont donnés respectivement dans la figure 3.5 et la figure 3.6 (en bas).
- **La forme d'onde** de la parole œsophagienne comme illustré dans la figure 3.5 (en haut) est une représentation bidimensionnelle, où le temps est représenté sur l'axe des abscisses et l'amplitude sur l'axe des ordonnées. Elle est utilisée pour pouvoir

détecter les silences, les courtes pauses, les bruits et les régions périodiques correspondant aux phonèmes ainsi que la transition qui permet le passage entre deux phonèmes successifs. Nous pouvons observer la différence et le bruit du signal œsophagien en comparant la forme d'onde du signal de la parole normale pour la même phrase représentée dans la figure 3.6 (en haut).

- **L'analyse** des différents paramètres prosodiques tels que l'intensité, la fréquence fondamentale, l'énergie et les formant aident et rendent objectif ce qui échappe parfois au yeux et à l'oreille humaine.

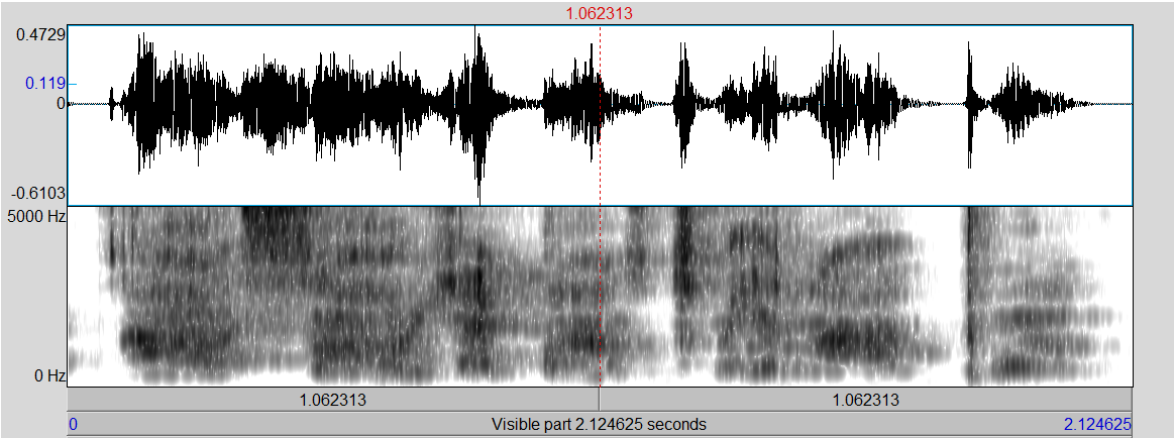

FIGURE 3.5: *Spectrogramme (en bas) et forme d'onde (en haut) du signal de la parole œsophagienne pour la phrase : "On songe à construire un pont"*

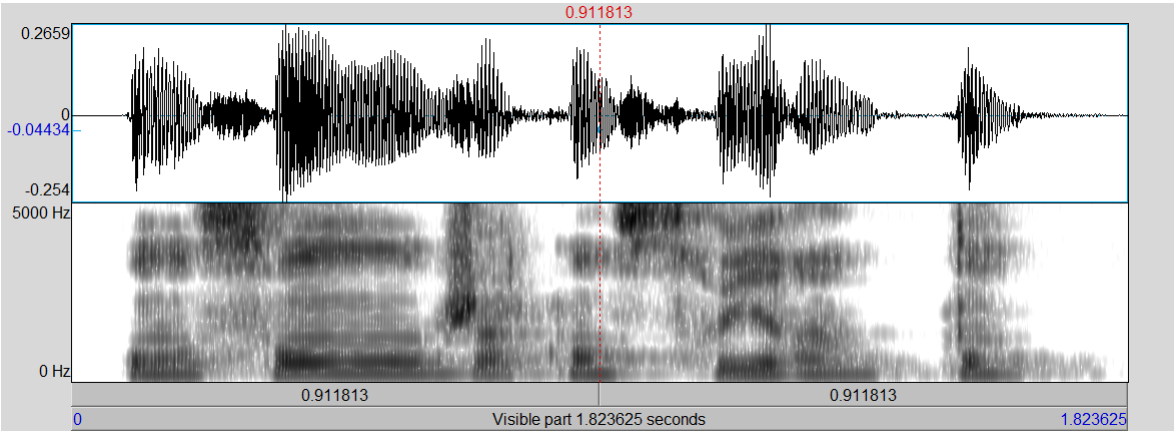

FIGURE 3.6: *Spectrogramme (en bas) et forme d'onde (en haut) du signal de la parole laryngée pour la phrase : "On songe à construire un pont"*

Tous ces critères de segmentation visuelle ne pourront pas bien sûr remplacer l'écoute du signal de la parole (oreille humaine). Effectivement, ce n'est pas facile de déceler précisément les frontières entre deux phonèmes successifs. En effet, prendre une décision sur

l'emplacement final d'une frontière s'avère souvent d'une grande subjectivité. Pour cette raison, le signal de la parole doit être écouté et analysé à plusieurs reprises.

Il est important d'utiliser des moyens matériels et logiciels d'aide à la segmentation afin de réduire le temps et l'effort humain nécessaire. Ces outils ont pour but de faciliter la tâche que ce soit pour l'étiquetage, la segmentation manuelle, ou pour la vérification et la correction de ces derniers. Il existe plusieurs logiciels permettant de visualiser le spectrogramme et la forme d'onde d'un signal de la parole, et d'éditer et d'aligner les transcriptions orthographiques et phonétiques sur ce signal, tels que Praat[4] , Wavesurfer[5] , SFS[6] ,WinSnoori[7].

Dans notre étude, nous avons choisi le logiciel Praat parce qu'il permet l'analyse des données acoustiques en calculant les paramètres prosodiques telles que l'intensité, la fréquence fondamentale ainsi que d'autres paramètres tels que l'énergie et les formants. Cet outil permet de segmenter le fichier audio en mots et en phonèmes en ajoutant manuellement des frontières et en étiquetant chaque intervalle (l'espace entre les deux frontières assignées). L'étiquetage est stocké dans un fichier TextGrid, qui a une structure particulière qui indique le temps de début et de fin pour chaque étiquette ainsi que la lecture vocale de ce segment. Un exemple d'une segmentation manuelle en mots et en phonèmes en utilisant le logiciel Praat pour la phrase : "On songe à construire un pont" est donné dans la figure 3.7. Les lignes verticales en bleu représentent les frontières entre les segments. Les formants sont représentés par des lignes pointillées en rouge sur la zone du spectrogramme. Le contour intonatif de la F0 est affiché en bleu et la courbe de l'intensité est tracée par une ligne jaune.

Comme mentionné précédemment, la détection d'une transition d'un phonème à un autre est un processus très délicat. En effet, le signal de la parole d'une phrase donnée n'est pas constitué de segments visuellement délimités. La difficulté de la segmentation manuelle se pose entre et à l'intérieur des mots. Cet inconvénient est facilement observé dans la figure précédente (voir figure 3.7), en regardant la forme d'onde sur la totalité de

4. http://www.fon.hum.uva.nl/praat/
5. http://www.speech.kth.se/wavesurfer/
6. https://www.phon.ucl.ac.uk/resource/sfs/
7. http://www.loria.fr/~laprie/WinSnoori/

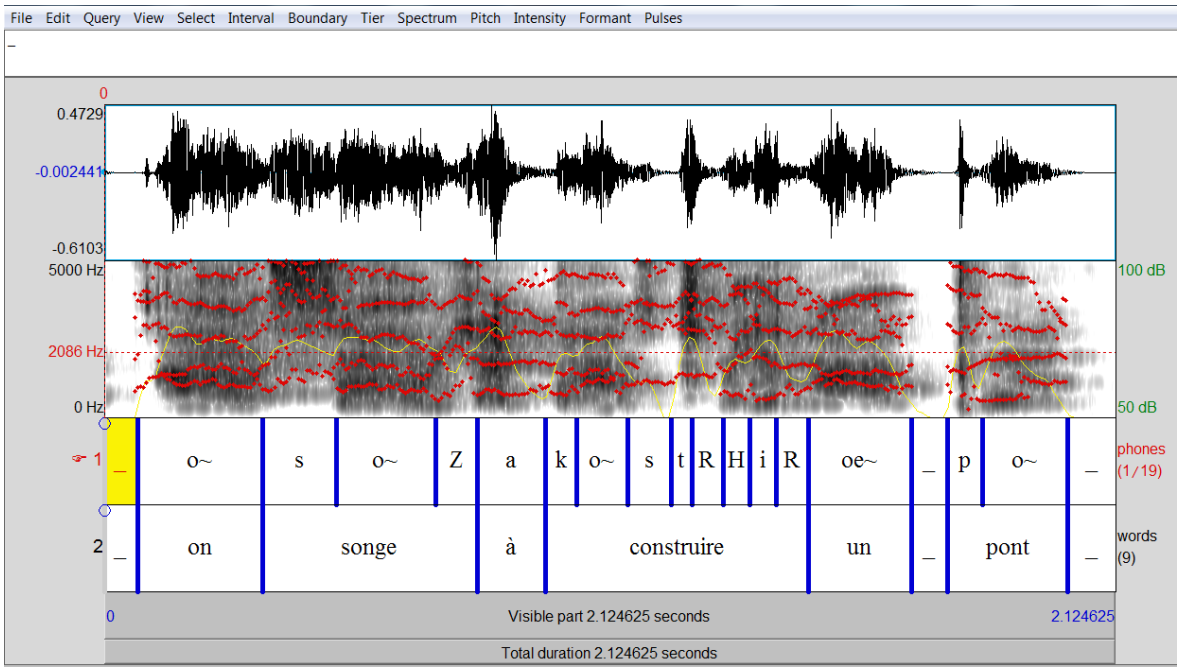

FIGURE 3.7: *Segmentation manuelle en mots et en phonèmes en utilisant le logiciel Praat pour la phrase : "On songe à construire un pont".*

la phrase à segmenter. Heureusement, le logiciel Praat permet aussi de faire des zooms sur les segments (mots, phonèmes, intervalles de temps). La fonction zoom, qui présuppose que l'on a sélectionné un segment sonore, est indiqué par un cadre rose. La figure 3.8 illustre le zoom de la segmentation manuelle du mot "songe" de la phrase précédemment décomposée. Cette fonction, permet de mieux déceler les marques de séparation entre deux phonèmes et aussi entre les mots et ainsi de faciliter la tâche de segmentation.

L'étiquetage phonétique des phrases a été réalisé avec SAMPA (Speech assessment Methods Phonetic Alphabet). Cette méthode d'étiquetage offre l'avantage d'utiliser uniquement des caractères ASCII simple. Avec SAMPA, on peut utiliser jusqu'à deux caractères pour représenter un phonème. Il existe une autre méthode standard de transcription phonétique appelé l'Alphabet Phonétique International (API). Malheureusement, dans la méthode API, chaque phonème est représenté par un symbole qui peut ne pas être saisi sur un clavier d'ordinateur. Le tableau 3.1, décrit la liste des 36 étiquettes phonétiques de la langue française utilisées dans notre base de données FPSD, avec la correspondance API, SAMPA et des exemples.

Numéro	API	SAMPA	Exemple	Numéro	API	SAMPA	Exemple
1	p	p	**p**ont [po~]	19	j	j	**i**on [jo~]
2	b	b	**b**on [bo~]	20	m	m	**m**ont [mo~]
3	t	t	**t**emps [ta~]	21	n	n	**n**om [no~]
4	d	d	**d**ans [da~]	22	ŋ	N	ri**ng** [riN]
5	k	k	**c**oût [ku]	23	l	l	**l**ong [lo~]
6	g	g	**g**ant [ga~]	24	ʁ	R	**r**ond [Ro~]
7	f	f	**f**emme [fam]	25	w	w	q**u**oi [kwa]
8	v	v	**v**ent [va~]	26	ɥ	H	j**u**in [ZHe~]
9	s	s	**s**ans [sa~]	27	i	i	s**i** [si]
10	z	z	**z**one [zOn]	28	e	e	bl**é** [ble]
11	ʃ	S	**ch**amp [Sa~]	29	ɛ	E	s**ei**ze [sEz]
12	ʒ	Z	**g**ens [Za~]	30	a	a	p**a**tte [pat]
13	ɔ	O	c**o**mme [kOm]	31	ø	2	d**eu**x [d2]
14	o	o	gr**o**s [gRo]	32	œ	9	n**eu**f [n9f]
15	u	u	d**ou**x [du]	33	oẽ	9~	br**un** [br9~]
16	y	y	d**u** [dy]	34	ẽ	e~	v**in** [ve~]
17	ə	@	d**e** [d@]	35	ã	a~	v**en**t [va~]
18	sil	- ou sil	**silence**	36	õ	o~	b**on** [bo~]

TABLEAU 3.1: *La transcription SAMPA des phonèmes français standards*

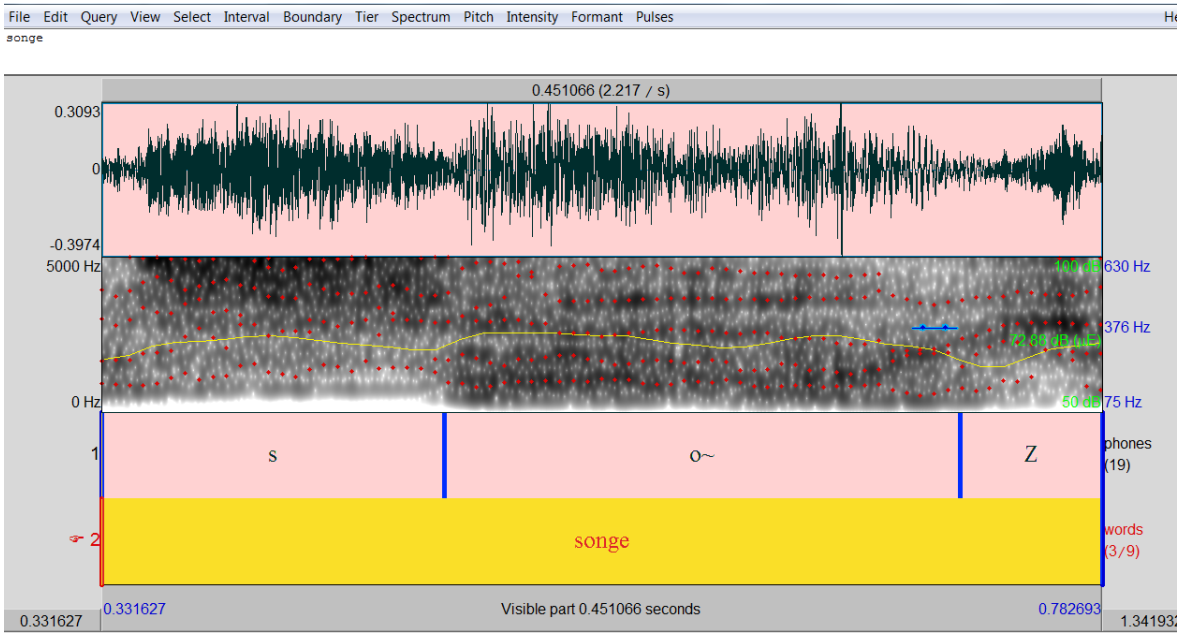

FIGURE 3.8: *Zoom du mot : "songe", sur le signal de la parole pour la phrase précèdemment segmentée : "On songe à construire un pont"*

3.4 Système de reconnaissance automatique de la parole œsophagienne

La reconnaissance et l'évaluation de la parole œsophagienne a toujours été la préoccupation clinique principale dans le domaine biomédical et la technologie de la parole [DIBAZAR et collab., 2006; PRAVENA et collab., 2012]. En général l'évaluation est effectuée par une variété de mesures se concentrant sur le signal et l'intelligibilité de la parole (comment un signal de la parole est entendu par d'autres). Elle est souvent associée à un jugement de perception. Cette méthode subjective incontournable consiste à évaluer la qualité vocale et décoder la parole par une simple écoute attentive. Cependant, l'analyse perceptuelle de la parole œsophagienne est longue et coûteuse car elle est sujette à diverses imperfections. En effet deux auditeurs non expérimentés peuvent fournir des jugements de perception différents sur le même signal acoustique (variabilité inter-auditeurs). En outre, des jugements variables dans le temps peuvent être fournis par un auditeur sur le même signal acoustique (variabilité intra-auditeur). D'autres personnes pourront ne pas être en mesure de comprendre ou décoder cette parole œsophagienne. Alors le recours à des jury d'experts peut être nécessaire afin d'augmenter la fiabilité de cette analyse perceptuelle, ce qui s'avère finalement coûteux en temps et en ressources humaines et ne peut pas être planifié régulièrement. De nos jours, l'analyse instrumentale dite "objecti-

ve" [WUYTS et collab., 2000; YU et collab., 2001] est de plus en plus utilisée. Elle s'appuie sur l'analyse de mesures acoustiques, aérodynamiques et physiologiques. Ces mesures peuvent être directement extraites du signal de la parole en utilisant un système informatique. Tout comme l'évaluation subjective de perception, les méthodes objectives comportent aussi des limites. Par exemple, l'analyse instrumental peut être très dépendantes de la population de patients examinés en matières de qualité et de quantité. En outre, se servir des appareils spécifiques de mesure peut s'avérer coûteux, ne permettant pas ainsi d'appliquer couramment cette technique.

Ces restrictions nous ont conduit récemment dans [LACHHAB et collab., 2014], à proposer une technique rapide et simple pour décoder et évaluer la parole œsophagienne (pathologique) en appliquant un système RAP continue sur notre propre base de données FPSD décrite dans la section 3.3. Le système de reconnaissance conçu pour cette tâche particulière, a été mis en œuvre à l'aide de la plate-forme HTK [YOUNG et collab., 2006], en utilisant des modèles HMM/GMM monophones (contexte-independant). Les vecteurs acoustiques sont transformés linéairement par la méthode HLDA [KUMAR et ANDREOU, 1998] détaillée dans la section 2.8.2 afin de réduire leur dimension dans un espace restreint qui augmente l'information discriminante. Dans les sous-sections qui suivent, nous décrirons la mise œuvre de notre système de reconnaissance automatique de la parole œsophagienne.

3.4.1 Pré-traitement des données acoustiques

Le système de reconnaissance de phonèmes utilise les Mel-Frequency Cepstral Coefficients MFCC [DAVIS et MERMELSTEIN, 1980] et l'énergie, ainsi que les coefficients différentiels de ces paramètres. Le signal est échantillonné à 16 KHz et pré-accentué avec un facteur de 0.97. Les 12 premiers coefficients cepstraux (c_1 à c_{12}) sont concaténés avec le logarithme de l'énergie de la trame pour former des vecteurs MFCC statiques de 13 coefficients (12MFCC + E). Ces coefficients sont calculés en utilisant une fenêtre de Hamming de 25 ms décalée toutes les 10 ms et à l'aide d'un banc de 26 filtres dans une échelle de fréquence Mel. Nous avons aussi inclus les coefficients différentiels d'ordre 1,2 et 3 appelés coefficients dynamiques ($\Delta, \Delta\Delta$ et $\Delta\Delta\Delta$) automatiquement en utilisant la paramétrisation de l'outil HTK. Nous travaillons donc initialement avec des vecteurs ayant au plus d=52 coefficients. Ensuite cet espace de dimension d est réduit par la méthode

HLDA appliquée sur tous les vecteurs (apprentissage et test) pour avoir des vecteurs plus pertinents et plus discriminants avec 39 coefficients (d=39) qui représentent la dimensionnalité de référence utilisée dans la majorité des systèmes RAP.

3.4.2 Apprentissage du système de reconnaissance automatique de la parole œsophagienne

Notre système de reconnaissance de la parole œsophagienne utilise comme unités acoustiques les 36 phonèmes de la transcription française SAMPA représentée dans le tableau 3.1 (dans la sous section 3.3). Ces phonèmes sont tous modélisés par la topologie classique HMM/GMM gauche-droite à 5 états. L'état initial et l'état final ont la particularité de ne pas émettre d'observation, mais de servir uniquement à la connexion des modèles en parole continue (seulement les 3 états intermédiaire sont émetteurs). L'apprentissage des modèles est le point de départ de tous les systèmes RAP et certainement le plus crucial. Il consiste à déterminer les paramètres optimaux $\{A, \pi_i, B\}$ (voir la section 2.5 : reconnaissance parole normale). Notre système RAP est implémenté à partir de la plate-forme HTK. Pour chaque modèle phonétique HMM, l'outil HInit initialise les probabilités d'émission des observations et de transitions d'états à l'aide de la méthode itérative des "k-moyennes segmentales" basée sur l'algorithme de Viterbi. Ces paramètres sont affinés par une estimation MLE effectuée par l'algorithme de Baum-Welch [BAUM, 1972] en utilisant l'outil HRest. La phase finale de l'apprentissage consiste à ré-estimer simultanément l'ensemble des modèles sur la parole continue grâce à l'outil HERest.

Il est important de choisir le nombre nécessaire de gaussiennes attribuées à chaque état en réalisant le meilleur compromis entre une bonne modélisation des HMMs monophones et le nombre limité de données d'apprentissage. Un mauvais apprentissage peut être observé, lors de l'utilisation d'un nombre élevé de gaussiennes, dû à la quantité de données d'apprentissage disponible, car le nombre d'échantillons pour chaque phonème est limité. Dans notre cas on a utilisé 16 gaussiennes par état sauf pour le phonème /N, qui ne peut avoir un nombre de gaussiennes supérieur à 14 par état.

3.4.3 Décodage de la parole œsophagienne

Le décodage de phonèmes est un processus délicat car on ne connaît pas la segmentation des phrases de test en phonèmes. En outre les modèles HMMs monophones supposent que la parole est produite comme une concaténation de phonèmes qui ne sont pas affectés par les contextes phonétiques gauche/droite et droite/gauche (contexte indépendant). Pour effectuer la reconnaissance (décodage), il est essentiel d'identifier la séquence d'états qui a généré les observations données. En fait, à l'aide de cette séquence d'états, nous pouvons facilement trouver la chaîne de phonèmes la plus probable qui correspond aux paramètres observés. Cette tâche est réalisée grâce à l'algorithme de décodage Viterbi appliqué sur chacune des phrases de test de notre corpus FPSD en se servant des paramètres optimaux $\{A, \pi_i, B\}$ déjà estimés. Ce décodage est amélioré par l'inclusion d'un modèle de langage bigramme, calculé sur la partie entière d'apprentissage de notre base de données FPSD. Ce langage bigramme a été construit statistiquement en utilisant seulement les 425 phrases à partir des modules HTK. Certes, la quantité de phrases disponible est insuffisante pour une parfaite estimation des probabilités d'occurrence de deux phonèmes successifs. Néanmoins, malgré cette faible quantité de phrases un gain d'environ 10% concernant le taux de reconnaissance phonétique (Accuracy) a été observé dans les résultats. En plus, ce modèle bigramme peut être bien sûr enrichi par divers contenus textuels issus de grandes bases de données françaises en vue d'améliorer les performances de notre système.

3.4.4 Expériences et résultats

Notre système de reconnaissance de la parole œsophagienne a été évalué sur notre corpus FPSD à l'aide des 36 étiquettes phonétiques SAMPA (voir tableau 3.1).

Nous avons effectué 4 séries d'expériences sur notre système RAP utilisant la voix œsophagienne pour évaluer l'apport des coefficients différentiels et de la transformation HLDA. Dans la première expérience nous avons travaillé avec des vecteurs de dimension d=39 (12 MFCC, E; 12 ΔMFCC, ΔE; 12 $\Delta\Delta$MFCC, $\Delta\Delta$E) qui représentent le cas de référence dans la plupart des systèmes RAP. Pour la deuxième expérience la dérivée d'ordre 3 ($\Delta\Delta\Delta$) est incluse dans l'espace des vecteurs afin d'augmenter leur dimension à d=52 (12 MFCC, E; 12 ΔMFCC, ΔE; 12 $\Delta\Delta$MFCC, $\Delta\Delta$E; 12 $\Delta\Delta\Delta$MFCC, $\Delta\Delta\Delta$E). La troisième expérience consiste à appliquer la transformation discriminante HLDA (39→39) sur les

39 coefficients utilisés dans l'expérience 1 sans réduction de dimensionnalité. Tandis que dans la quatrième et dernière expérience la dimentionnalité de 52 (coefficients) utilisée dans l'expérience 2 à été réduite à 39 (coefficients) grâce à la transformation HLDA (52→39).

Le tableau 3.2 présente les résultats de décodage obtenus pour les 4 expériences décrites ci-dessus sur la partie test de notre base de données FPSD de la parole œsophagienne.

36 HMMs monophone avec 16 Gaussiennes par état + Bigramme	**Accuracy (%)**	**Correct (%)**
Expérience 1 : 39 coefficients MFCC	61.89	67.62
Expérience 2 : 52 coefficients MFCC	58.49	65.29
Expérience 3 : HLDA coefficients (39 → 39)	62.31	66.88
Expérience 4 : HLDA coefficients (52 → 39)	**63.59**	**69.43**

TABLEAU 3.2: *L'apport des coefficients différentiels et de la transformation HLDA sur le taux de reconnaissance phonétique (Accuracy) obtenu sur la partie Test de notre base de données FPSD*

Nous remarquons d'après les résultats observés dans la quatrième expérience (4), que le taux de reconnaissance (Accuracy) est amélioré significativement par rapport aux autres expériences..

3.5 Conclusion

Notre système de reconnaissance automatique de la parole œsophagienne, basé sur des modèles HMM/GMM monophones (indépendants du contexte) a apporté une amélioration significative du taux de reconnaissance le fixant à 63.59% grâce à la transformation discriminante HLDA et l'introduction de coefficients différentiels d'ordre élevé. Les performances de notre système de reconnaissance sont encourageants. Certainement, ces résultats peuvent encore être améliorés par l'extension de notre corpus FPSD afin de rendre possible l'utilisation des modèles HMM dépendants du contexte (triphones) et aussi en employant un modèle de langage bigramme plus précis.

Chapitre 4

Amélioration de la reconnaissance de la parole alaryngée

« *La nature fait les hommes semblables, la vie les rend différents.* »

Confucius

4.1 Les recherches antérieures et actuelles sur l'amélioration de la parole alaryngée

La parole alaryngée (pathologique) se caractérise par une perturbation de bruit élevé, une faible intelligibilité et une fréquence fondamentale instable. Ces caractéristiques qui sont très différentes de celles de la parole laryngée (normale) produisent une voix rauque, grinçante et non naturelle, difficile à comprendre. Pour cette raison, diverses méthodes ont été proposées pour améliorer la qualité et l'intelligibilité de la parole alaryngée. L'objectif principal de ces travaux et le rétablissement des caractéristiques de la voix laryngée dans la mesure du possible. [YINGYOUNG, 1990] a proposé d'améliorer la qualité des voyelles pour la voix trachéo-œsophagienne à l'aide d'un codage par prédiction linéaire (LPC). Les fonctions d'erreurs de prédiction normalisées ont été utilisées pour choisir les paramètres de contrôle de l'analyse. Les trames dont les erreurs de prédiction normalisées étaient proches d'un minimum ont été utilisées pour sélectionner les pôles de la fonction de transfert du conduit vocal. Cette fonction de transfert a permis de synthétiser les voyelles. La nouvelle entrée excitative est basée sur une impulsion glottale naturelle.

De son côté, [MATUI et collab., 1999] ont proposé d'améliorer les caractéristiques spectrales de la voix œsophagienne en se basant sur la technique de synthèse par formants. Le remplacement du voisement humain par des signaux d'excitation artificiels constitue une approche alternative. Dans [LOSCOS et BONADA, 2006], un contour de pitch artificiel a été créé à partir de l'enveloppe de l'énergie de la parole pour remédier au problème d'instabilité de la fréquence fondamentale F0. Les auteurs de l'étude [ALI et JEBARA, 2006], ont proposé de modifier la voix d'un locuteur alaryngée par le déplacement des fréquences des formants vers une bande plus haute étant donné que la longueur du conduit vocal a été raccourcie. [DEL POZO et YOUNG, 2006], utilisent une forme d'onde glottale synthétique combinée avec un modèle de réduction du jitter et shimmer pour réduire le bruit et le grincement de la parole trachéo-œsophagienne originale. Le jitter mesure le niveau de perturbation de la fréquence fondamentale F0, donc la déficience de vibration des cordes vocales de l'appareil phonatoire. Tandis que le shimmer mesure le niveau de perturbation de l'intensité vocale, perturbation liée au passage brusque et anormal d'une voix forte vers une voix faible. [TÜRKMEN et KARSLIGIL, 2008] ont proposé la méthode MELP (Mixed-Excitation Linear Prediction), qui consiste à synthétiser une parole nor-

male en utilisant l'estimation du pitch et la correction des formants pour les phonèmes voisés de la voix chuchotée. Les phonèmes non voisés, ne sont pas modifiés dans cette approche. Cependant, cette technique ne convient pas à un fonctionnement en temps réel. Un autre exemple a été rapporté par [SHARIFZADEH et collab., 2010], nommé CELP (Code-Excitation Linear Prediction). Celle-ci tente de produire des caractéristiques plus naturelles par la reconstruction des éléments manquants liés au pitch pour la parole chuchotée. Cependant, il est encore très difficile de générer des signaux d'excitation réalistes similaires à ceux naturellement générés par les vibrations des cordes vocales.

D'autres tentatives pour la correction ou l'amélioration de la parole alaryngée en se basant sur la modification des caractéristiques acoustiques ont été proposées : elles sont fondées sur la réduction du bruit de fond basé sur le masquage auditif [LIU et collab., 2006] ; la réduction du bruit de respiration généré par l'effet du passage de l'air via le conduit vocal sans constrictions, combinée avec une stabilisation des pôles du système modélisant ce conduit à l'aide des paramètres LPC [GARCIA et collab., 2002, 2005] ; le filtrage en peigne [HISADA et SAWADA, 2002] ; le débruitage de la parole électrolarynx par soustraction spectrale [COLE et collab., 1997]. Cette dernière méthode de type soustractive est limitée et manque de précision dans l'estimation du bruit de fond. De son côté, [MANTILLA-CAEIROS et collab., 2010] a proposé de remplacer les segments sonores voisés de la parole œsophagienne, sélectionnés (à l'aide de techniques de reconnaissance de formes) par les segments sonores de la parole normale correspondante. Le silence et les segments non voisés ne subissent aucun changement. Un autre travail rapporté dans [DEL POZO et YOUNG, 2008], consiste à corriger les durées des phonèmes de la parole trachéo-œsophagienne par celles prédites en utilisant des arbres de régression construits à partir des données de la parole laryngée.

Les techniques dites de "conversion de la voix" ont été proposées afin de rapprocher les caractéristiques de la voix pathologique vers celles de la parole laryngée. La conversion vocale est souvent utilisée pour la synthèse vocale. Généralement, la conversion a été employée afin de transformer la voix d'un locuteur source en celle d'un locuteur cible (laryngée). Cette méthode est basée sur l'apprentissage d'une "fonction de conversion", qui s'obtient en modélisant les densités de probabilités conjointes des paramètres cepstraux des voix source et cible.

L'un des premiers systèmes pour améliorer la parole alaryngée en se basant sur la conversion vocale a été proposé par [NING et YINGYONG, 1997]. Ce système utilise la quantification vectorielles (QV) et la Régression Linéaire Multivariée (RLM) pour l'estimation de la fonction de conversion. La QV a été modifiée par un chirp transformé en Z (généralisation de la transformée de fourrier discrète), qui subit ensuite une pondération cepstrale afin de diminuer la bande passante des formants. Ce système a été appliqué à la parole alaryngée et a été évalué par des tests de perception. Les expériences effectuées ont indiqué que les auditeurs préfèrent la parole alaryngée convertie par rapport à l'originale (alaryngée non convertie). Récemment dans [DOI et collab., 2014], la qualité et l'intelligibilité de la parole alaryngée a été améliorée par l'approche de conversion vocale "EigenVoice". La parole alaryngée convertie a été re-synthétisée (reconstruite) afin d'évaluer sa qualité. Pour tenir compte des différentes caractéristiques du locuteur cible et pour palier le manque de données (peu de phrase pour faire l'apprentissage), cette méthode propose d'ajuster les vecteurs moyens par des poids de pondération appris durant la phase d'apprentissage. En complément à cette technique, [TANAKA et collab., 2014] intègre dans un nouveau système hybride la méthode de réduction de bruit par la soustraction cepstrale [BOLL, 1979] et en utilisant la conversion de la voix statistique afin de prédire les paramètres d'excitation. Ces deux approches récentes visent à améliorer l'estimation des caractéristiques acoustiques afin de reconstruire un signal converti avec une meilleure intelligibilité. Cependant, le processus de conversion utilisé dans ces deux méthodes est trop complexe et peut générer des erreurs dans l'estimation des paramètres (beaucoup d'informations nécessaires à la génération du signal sont perdues) et donc créer des segments de sons non naturels en raison d'un manque crucial de signaux d'excitation réalistes liés aux paramètres spectraux convertis. Par conséquent, dans la pratique, il est difficile de compenser les différences existantes au niveau des paramètres acoustiques alaryngés par rapport à celles de la parole laryngée.

Pour ces raisons, nous proposons dans [LACHHAB et collab., 2015], un système hybride basé sur un algorithme de conversion statistique GMM de la voix pour améliorer la reconnaissance de la parole œsophagienne. Ce système hybride vise à compenser les distorsions présentes dans les vecteurs acoustiques de la parole œsophagienne à l'aide d'un procédé de conversion de la voix. La parole œsophagienne est convertie en parole laryngée "cible" à l'aide d'une fonction de transformation estimée statistiquement d'une

façon itérative. Nous n'avons pas appliqué un module de re-synthèse vocal pour reconstruire le signal de la parole convertie, vu que notre système de reconnaissance automatique de la parole utilise directement les vecteurs Mel cepstraux convertis comme paramètres d'entrée. En outre, les vecteurs acoustiques sont linéairement transformés par la méthode HLDA (analyse discriminante linéaire hétéroscédastique) pour réduire leur dimension dans un espace restreint ayant de bonnes propriétés discriminantes. Les résultats expérimentaux démontrent que le système proposé fournit une amélioration du taux de reconnaissance de phonèmes (Accuracy) avec une augmentation absolue de 3.40% par rapport au système de base, sans transformation HLDA ni conversion de voix.

4.2 Principes d'un système de conversion de la voix

La conversion vocale est un processus qui consiste à transformer le signal de la parole d'un locuteur source, de façon à ce qu'il semble à l'écoute, avoir été prononcé par un locuteur cible. En d'autre termes, la modification est effectuée seulement sur les caractéristiques du signal de la parole dépendantes du locuteur, tels que la forme spectrale, les formants, la fréquence fondamentale (F0), l'intonation et l'intensité afin de changer l'identité du locuteur, sans pour autant perdre l'information ou modifier le contenu de la phrase prononcée.

Cette technologie a plusieurs domaine d'applications, nous pouvons citer : la synthèse de la parole personnalisée à partir de texte TTS (Text-To-Speech) [KAIN et MACON, 1998; STYLIANOU et collab., 1998], la conversion vocale en général [EN-NAJJARY, 2005], l'amélioration ou la correction de la voix alaryngée [DOI et collab., 2014; NAKAMURA et collab., 2012; NING et YINGYONG, 1997; TANAKA et collab., 2014; TODA et collab., 2009]. Les systèmes de conversion de voix adoptent tous une structure similaire qui est résumée dans la figure 4.1.

Ils se décomposent en deux phases principales :

- **Une phase d'apprentissage** durant laquelle les phrases prononcées par les locuteurs source et cible, subissent une étape de paramétrisation (analyse acoustique). Une séquence de vecteurs acoustiques est extraite des ondes sonores correspondantes. Ces données d'apprentissage des locuteurs, source et cible passent par une

étape d'alignement des vecteurs (trames). Cette étape consiste à associer chaque vecteur source à son vecteur cible correspondant. Cette correspondance est aisément obtenue si on dispose de corpus parallèles de voix source et cible qui contiennent des phrases possédant le même contenu phonétique. Cet alignement est réalisé grâce à l'algorithme DTW (Dynamic Time Warping) [SAKOE et CHIBA, 1971] qui permet d'apparier deux à deux les vecteurs source et cible. La fonction de conversion optimale est estimée à partir de cette base de données alignée. Les vecteurs du locuteur source sont convertis en vecteurs du locuteur cible tout en minimisant l'erreur quadratique moyenne entre les vecteurs convertis et les vecteurs cible. Dans la littérature, diverses méthodes statistiques ont été proposées pour estimer la fonction de conversion : la quantification vectorielle [ABE et collab., 1988], la régression linéaire multivariée [NING et YINGYONG, 1997; VALBRET et collab., 1992], la déformation fréquentielle dynamique DFW (Dynamic Frequency Warping) [VALBRET et collab., 1992], les GMMs par l'estimation de l'erreur quadratique [STYLIANOU et collab., 1998] ou l'estimation de la probabilité conjointe source/cible [KAIN et MACON, 1998; TODA et collab., 2007; WERGHI et collab., 2010]. Certaines de ces méthodes seront détaillées dans les sections suivantes.

⊛ **Une phase de conversion** qui consiste à transformer trame par trame, les paramètres acoustiques issus du locuteur source vers leurs correspondants cible, en utilisant la fonction de conversion précédemment estimée. Un synthétiseur vocale est appliqué pour reconstruire le signal de la parole converti.

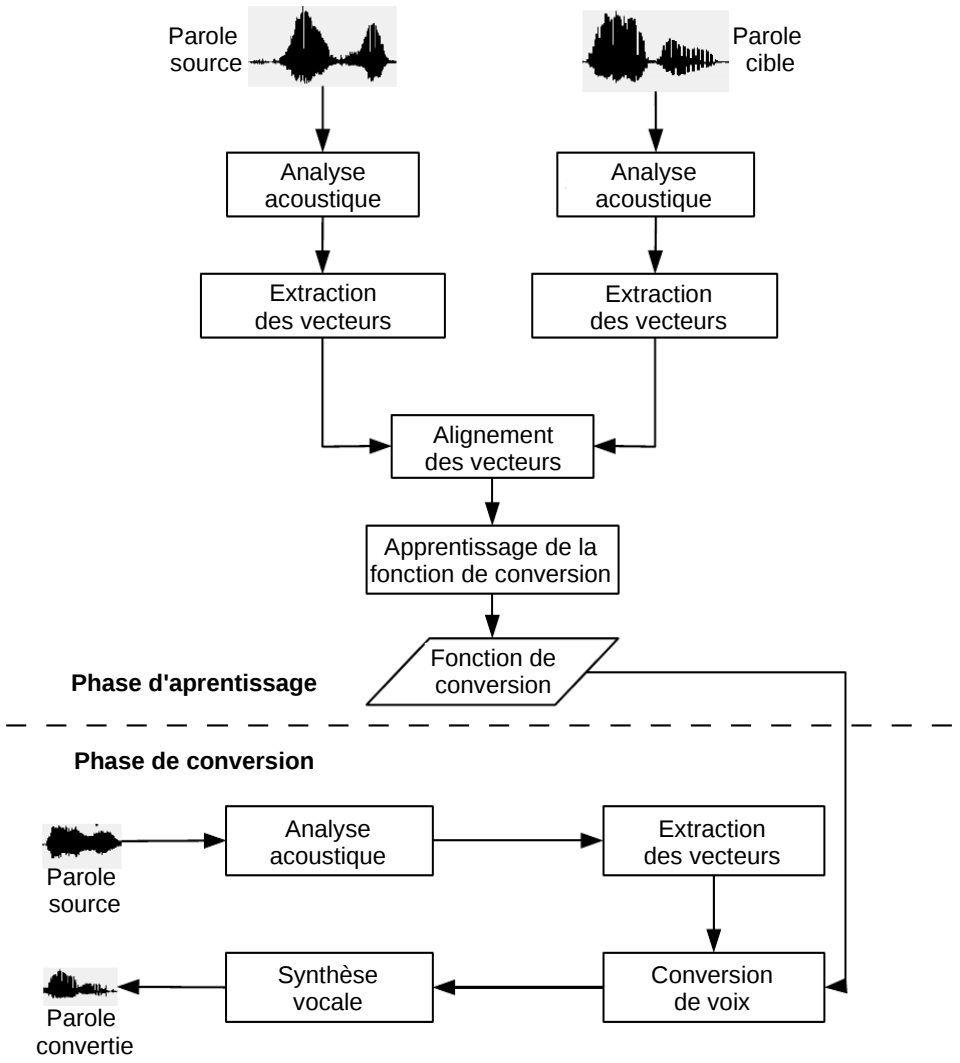

FIGURE 4.1: *Phases d'apprentissage et de transformation d'un système de conversion de voix.*

4.2.1 Analyse et paramétrisation

La nature des paramètres acoustiques utilisés dépend du système de conversion. Généralement, les plus utilisés dans le processus de conversion sont : CF (cepstres de Fourier), MFCC, LPC, LSF (Linear Spectral Frequency), HNM (Harminic Noise Model) ou par des paramètres relatifs aux formants. L'objectif de ces représentations est de réduire la dimensionnalité élevé de l'enveloppe spectrale correspondante au spectre d'amplitude du filtre modélisant le conduit vocal et le spectre de la source glottique.

4.2.2 L'alignement parallèle

L'alignement parallèle est utilisé lorsqu'on dispose de deux corpus parallèles de voix source et cible, dont les phrases prononcées possèdent le même contenu phonétique.

En général, l'alignement par DTW est la technique la plus utilisée dans les systèmes de conversion de voix. Elle consiste à trouver le chemin optimal qui met en correspondance les vecteurs acoustiques des locuteurs source et cible, c'est-à-dire à associer chaque vecteur source d'une séquence à un vecteur cible de l'autre séquence, en minimisant les coûts d'association. Le coût d'une association est calculé par la distance entre les deux vecteurs. La figure 4.2 représente un exemple d'alignement des vecteurs réalisé par l'algorithme DTW. Cette technique est applicable sur tout le signal de la parole [STYLIANOU et collab., 1998] [KAIN et MACON, 1998]

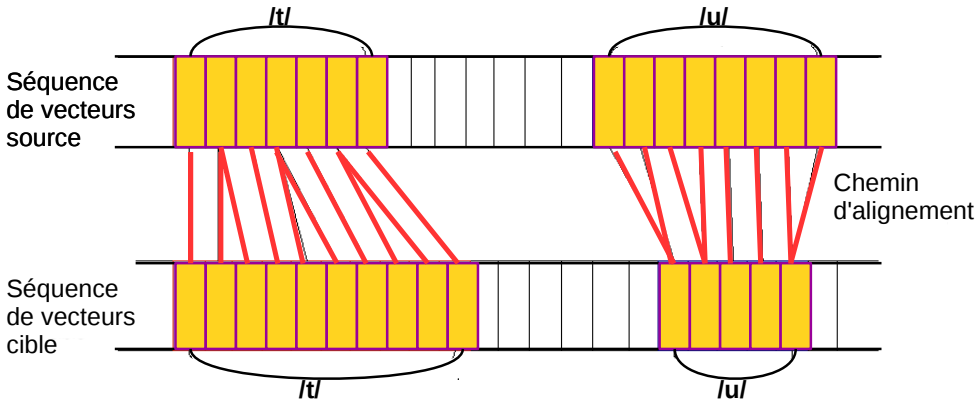

FIGURE 4.2: *Alignement temporel DTW entre les vecteurs source et cible.*

4.2.3 Apprentissage de la fonction de conversion

4.2.3.1 Conversion de voix par quantification vectorielle

La conversion de voix par quantification vectorielle est la première technique appliquée à la conversion de voix, proposée par [ABE et collab., 1988]. Le pitch, l'énergie et les paramètres spectraux sont considérés dans cette étude comme les paramètres acoustiques dépendants du locuteur. La quantification vectorielle consiste à projeter les paramètres acoustiques d'un espace de grande dimension, vers un espace de classes beaucoup plus réduit. Chaque classe est représentée par un vecteur particulier appelé "centroïde" (voir figure 4.3). Ce vecteur est lié à la distance minimale intra-classe.

La correspondance entre centroïdes source C_i^s et cible C_j^c, se fait par alignement DTW. Toutes les correspondances sont accumulées dans un histogramme qui agit en tant que

fonction de pondération. La correspondance des classes (dictionnaires) est déterminée par une combinaison linéaire des vecteurs du locuteur cible. Lors de la transformation, il suffit alors de remplacer chaque vecteurs cible par son homologue dans la liste de correspondance des dictionnaires créés. La parole est re-synthétisée grâce à ces nouveaux paramètres acoustiques convertis. Cette technique a l'avantage d'être simple et peu coûteuse en temps de calcul. Cependant, elle n'offre qu'une représentation discrète de la conversion.

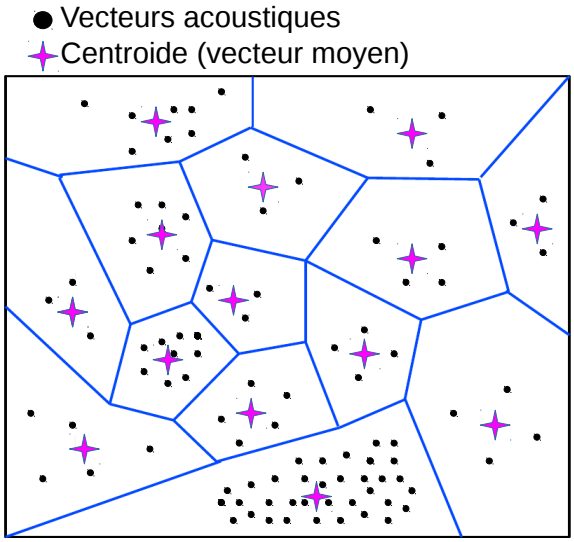

FIGURE 4.3: *Exemple d'une quantification vectorielle.*

4.2.3.2 Conversion de voix par réseaux de neurones multicouches

Un réseau de neurones multicouches (voir la figure 4.4), se compose d'une couche d'entrée qui reçoit les données de traitement, d'une ou plusieurs couches cachées (couches intermédiaires de traitement) et d'une couche de sortie. Chaque neurone est connecté à l'ensemble des neurones de la couche suivante, par des connexions dont les poids w_i jouent un rôle primordial dans l'apprentissage. La fonction de transformation des N vecteurs d'entrées x_i vers les M vecteurs de sorties y_i est définie par l'équation suivante :

$$\hat{y_i} = \mathrm{G}(\sum_{j=1}^{\mathrm{N}} w_{ij} x_i - \theta) \tag{4.1}$$

Avec :

- G : correspond à une fonction non linéaire du neurone
- θ : est un seuil ou biais.

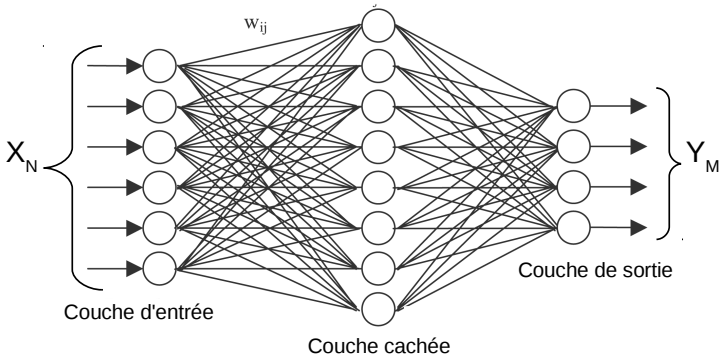

FIGURE 4.4: *Réseaux de neurones multicouches de N entrées et M sorties.*

L'algorithme d'apprentissage modifie, de façon itérative, les poids pour adapter la sortie obtenue $\hat{y}_i$ à la sortie désirée y_i. L'objectif est de chercher l'ensemble des poids $\mathscr{W}$, qui minimise l'erreur quadratique entre les sorties obtenues $\hat{y}_i$ et les sorties désirées y_i.

$$\mathscr{W} = \arg\min_{\mathscr{W}} \sum_{i=1}^{N} ||y_i - \hat{y}_i||^2 \tag{4.2}$$

Les poids du réseau de neurones sont ajustés grâce à la méthode d'apprentissage par rétro propagation du gradient de l'erreur [RUMELHART et collab., 1986]. Dans l'étude [NARENDRANATH et collab., 1995], la conversion par réseaux de neurones est utilisée sur les valeurs des trois formants comme entrée pour le locuteur source. Les sorties désirées sont les valeurs des trois formants issus du locuteurs cible. Tandis que dans [DESAI et collab., 2010], l'apprentissage de la fonction de conversion est exploitée sur les paramètres spectraux.

4.2.3.3 Conversion de voix par mélange de gaussiennes (GMM)

La conversion de voix par mélange de gaussiennes est une méthode standard initialement proposée par Styliannou [STYLIANOU et collab., 1998]. Le modèle GMM permet une modélisation probabiliste continue et efficace de l'espace acoustique d'un locuteur. Les discontinuités spectrales présentent dans tous les autres algorithmes de conversion de voix disparaissent et le naturel de la voix convertie est amélioré. Soit $X_N = [x_1, x_2, \ldots, x_N]$ la séquence de vecteurs acoustiques correspondant à la parole d'un locuteur source et $Y_N = [y_1, y_2, \ldots, y_N]$ la séquence de vecteurs acoustiques correspondante au même énoncé

prononcé par le locuteur cible. Supposons aussi que le nombre de vecteurs dans les deux séquences est égale à N.

La distribution de probabilité d'un vecteur x_n pour un modèle GMM à M composantes (gaussiennes) est définie par :

$$p(x_n) = \sum_{i=1}^{M} \alpha_i \mathcal{N}_i(x_n, \mu_i, \Sigma_i) \tag{4.3}$$

Chaque gaussienne est représentée par un vecteur moyen μ et une matrice de covariance Σ, α_i est le poids de pondération de la composante i, avec $\sum_{i=1}^{M} \alpha_i = 1, \alpha_i \geq 0$. L'algorithme EM [DEMPSTER et collab., 1977] est utilisé pour estimer les paramètres $(\alpha_i, \mu_i, \Sigma_i)$ du GMM. Une fois la classification par GMM effectuée, la fonction de conversion source→cible s'écrit comme une régression linéaire de la forme suivante :

$$\mathcal{F}(x_n) = \sum_{i=1}^{M} p(C_i|x_n)\left(\mu_i^y + \Sigma_i^{yx}\left(\Sigma_i^{xx}\right)^{-1}(x_n - \mu_i^x)\right) \tag{4.4}$$

Où $p(C_i|x_n)$ est la probabilité d'observer la classe C_i sachant x_n.

$$p(C_i|x_n) = \frac{\alpha_i \mathcal{N}(x_n, \mu_i^x, \Sigma_i^{xx})}{\sum_{j=1}^{M} \alpha_j \mathcal{N}(x_n, \mu_j^x, \Sigma_j^{xx})} \tag{4.5}$$

Le vecteur moyen μ_i^y et la matrice de covariance croisée Σ_i^{yx} de la gaussienne i sont estimés en minimisant la distance quadratique moyenne E entre les vecteurs convertis et les vecteurs cibles par la formule :

$$E = \min_{\mu^y, \Sigma^{yx}} \sum_{n=1}^{N} ||y_n - \mathcal{F}(x_n)||^2 \tag{4.6}$$

Où x_n et y_n désignent respectivement les vecteurs source et cible précédemment mis en correspondance par l'alignement DTW.

Dans [KAIN et MACON, 1998], l'auteur a amélioré la procédure d'apprentissage de la fonction de conversion en proposant, un modèle GMM conjoint qui dépend des paramètres source et cible (au lieu du modèle source proposé par [STYLIANOU et collab.,

1998]). Cette variante revient à estimer directement l'ensemble des paramètres à la fois source et cible $(\alpha_i, \mu_i^x, \mu_i^y, \Sigma_i^{xx}, \Sigma_i^{yx})$ de la fonction de conversion par l'algorithme EM. Cette approche rend l'estimation des paramètres source et cible plus stable numériquement. Les vecteurs correspondant source-cible sont concaténés conjointement dans un seul vecteur étendu, $\forall n \in [1,2,\ldots,N]$ on construit le vecteur $z_n = [x_n, y_n]'$ et ensuite on estime les paramètres GMM qui modélisent la densité de probabilité conjointe $p(z_n)$ suivante :

$$p(z_n) = p(x_n, y_n) = \sum_{i=1}^{M} \alpha_i \mathcal{N}_i(z_n, \mu_i, \Sigma_i) \tag{4.7}$$

$$\Sigma_i = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix} \quad \text{et} \quad \mu_i = \begin{bmatrix} \mu_i^x \\ \\ \mu_i^y \end{bmatrix}$$

L'étude faite par [TODA et collab., 2007], a tenté de trouver une solution au problème de sur-lissage de la conversion par GMM. La solution proposée consiste à calculer la variance globale effectuée sur les vecteurs convertis, et à maximiser la vraisemblance du modèle de conversion, en prenant en compte la trajectoire des vecteurs acoustiques. Bien que cette approche permette une réduction des erreurs dans la conversion la qualité de la voix convertie synthétisée est dégradée, car beaucoup d'informations nécessaires à la génération de la parole sont perdues dans cette approche de conversion.

4.3 La re-synthèse vocale

Différentes approches ont été proposées afin d'améliorer la qualité et l'intelligibilité de la parole chez les personnes laryngectomisées. La plupart de ces travaux [DOI et collab., 2014; NAKAMURA et collab., 2012; NING et YINGYONG, 1997; TANAKA et collab., 2014] utilisent un module de re-synthèse vocale afin de reconstruire le signal converti. Les modèles de re-synthèse vocale sont liés aux systèmes de conversion de la voix.

Les modèles de synthèse de la voix les plus couramment utilisés sont :

- **Le modèle PSOLA (Pitch-Synchronous Overlap-Add) :** est une technique basée sur la décomposition d'un signal de la parole en plusieurs segments qui se chevauchent [MOULINES et CHARPENTIER, 1990]. Chaque segment du signal analysé

représente une des périodes consécutives synchronisées sur le pitch, et l'addition-recouvrement de ces segments peut être utilisé pour la reconstruction du signal de la parole. PSOLA fonctionne directement sur la forme d'onde du signal, ce qui permet une synthèse de la parole sans perte de détails. Différentes variantes de la méthode PSOLA ont été proposées afin d'améliorer de façon significative la qualité de la parole synthétisée. Citons FD-PSOLA (Frequency Domain PSOLA) et TD-PSOLA (Time-Domain PSOLA) qui ont été utilisés dans différents travaux [TURK et ARSLAN, 2006; VALBRET et collab., 1992].

⊛ **Le Modèle Harmonique plus Bruit HNM (Harmonic Noise Model) :** est un modèle qui consiste à décomposer le signal de la parole S(t) en deux parties : une partie harmonique $h(t)$ et une partie bruitée $b(t)$ [STYLIANOU, 1996; STYLIANOU et collab., 1998]. La partie harmonique modélise la composante quasi-périodique des sons voisés du signal de la parole, tandis que la partie bruitée modélise la composante aléatoire du signal, comme le bruit de friction et les variations de l'excitation glottique d'une période à l'autre.

Le signal S(t) peut s'écrire ainsi :

$$S(t) = h(t) + b(t) \tag{4.8}$$

Avec :

$$h(t) = \sum_{n=0}^{N(t)} A_n(t) cos(2\pi t n f_0(t) + \phi_n(t) \tag{4.9}$$

Où $A_n(t)$,$\phi_n(t)$ correspondent à l'amplitude et la phase de la $n^{\text{ième}}$ harmonique à l'instant t. $f_0(t)$ est la fréquence fondamentale à l'instant t et N(t) correspond au nombre d'harmoniques inclus dans la partie harmonique à l'instant t.

En général, le signal sonore est caractérisé par des trames voisées et non voisées. Dans le cas des trames voisées, le spectre du signal est divisé en deux bandes délimitées (voir la figure 4.5) par la fréquence maximale de voisement f_m (fréquence de coupure). La bande inférieure du spectre (en dessous de la fréquence f_m) est représentée par la partie harmonique (signal passe-bas), tandis que la bande supérieure correspond à la partie bruitée (signal passe-haut).

Le modèle Auto Régressif (AR) variant dans le temps, permet de décrire le contenu fréquentiel de la partie bruitée représentant les trames non-voisées et le bruit de friction.

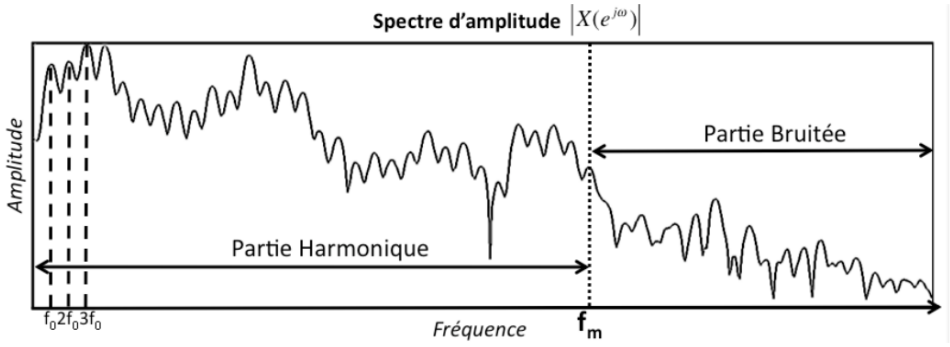

FIGURE 4.5: *Décomposition du spectre en bandes "harmonique" et "bruit" délimitées par la fréquence maximale de voisement f_m.*

Dans ce cas, la partie bruitée $b(t)$ est obtenue en filtrant un bruit blanc gaussien $u(t)$ par un filtre tout pôle $g(t)$ et en multipliant le résultat obtenu par une enveloppe d'énergie $e(t)$.

$$b(t) = e(t)[g(t) * u(t)] \tag{4.10}$$

La reconstruction du signal synthétique $\hat{S}(t)$ par la méthode HNM (Harmonique plus Bruit) est obtenu par l'addition de la partie harmonique $h(t)$ et de la partie bruitée $b(t)$.

$$\hat{S}(t) = h(t) + b(t) \tag{4.11}$$

⊛ **Le modèle STRAIGHT (Speech Transformation and Representation using Adaptative Interpolation of weiGHTed spectrum) :** est un modèle de synthèse vocale basé sur la théorie source-filtre [KAWAHARA, 1997; KAWAHARA et collab., 1999]. Ce modèle utilise trois composantes pour synthétiser la parole : a) la fréquence fondamentale F0 ; b) les coefficients d'apériodicité ; c) les coefficients spectraux. L'auteur de cette méthode propose une analyse spectrale (adaptative-pitch) combinée avec une reconstruction de surfaces en utilisant des fenètres adaptatives temps-fréquences. L'objectif de cette pocédure est d'obtenir une enveloppe spectrale dépourvue d'information due à la périodicité (élimination des effets de périodicité).

Le signal associé à un segment voisé est représenté comme la somme de K harmoniques comme suit :

$$s(t) = \sum_{k=1}^{K} \alpha_k(t) \sin[\int_{t_0}^{t} k(w(\tau) + w_k(\tau))d\tau + \phi_k] \tag{4.12}$$

Où $t_0 = 1/F0$ et $w(\tau)$ correspond à une fenêtre temporelle. ϕ_k, α_k et $w_k(\tau)$ correspondent respectivement à la phase, l'amplitude et la pulsation associée à la $k^{ième}$ harmonique.

$$w(\tau) = \frac{1}{\tau_0} e^{-\pi(\tau/\tau_0)^2} \tag{4.13}$$

Les coefficients d'apériodicité correspondent à l'énergie associée au fréquences non-harmoniques. Ces coefficients sont définis comme la normalisation des composantes de bruit (enveloppe spectrale supérieure) par les composantes périodiques du signal (enveloppe spectrale inférieure).

Ce modèle a été largement utilisé dans la conversion de la voix [DESAI et collab., 2010; DOI et collab., 2014; TANAKA et collab., 2014; TODA et collab., 2007].

4.4 Évaluation de la conversion de voix alaryngée

L'étape d'évaluation de la conversion de la voix alaryngée vers une voix normale est essentielle pour mesurer les progrès effectués, par exemple : évaluer la qualité, l'intelligibilité et le naturel de la parole convertie synthétisée. Il existe deux genres d'évaluation : objective et subjective. Généralement, les tests objectifs sont effectués par des mesures de distance entre les vecteurs acoustiques cible et convertie, tandis que les tests subjectifs dits aussi de perception sont basés sur l'évaluation auditive pour mesurer la qualité et l'intelligibilité de la voix convertie. Cependant aucune fonction de mesure objective ne permet à ce jour de remplacer totalement l'oreille humaine ou les tests de perception par un jury d'experts. Dans le but d'évaluer la parole œsophagienne, nous avons proposé dans [LACHHAB et collab., 2014] une simple et rapide technique en appliquant un système de reconnaissance automatique de la parole sur notre propre base de données FPSD. L'objectif est d'extraire une quantité conséquente de l'information phonétique contenue dans le signal de cette parole œsophagienne.

4.4.1 Évaluation objective

Parmi les tests objectifs proposés dans la littérature, on trouve :

- **L'erreur de distorsion normalisée :** permet de mesurer le rapprochement entre voix cible et convertie [ABE et collab., 1988]. L'évaluation consiste à calculer la distance spectrale DS entre les deux signaux de parole, par la relation suivante :

$$\mathrm{R} = \frac{\mathrm{DS}(convertie, cible)}{\mathrm{DS}(source, cible)} \tag{4.14}$$

- **Taux d'erreur des segments voisés/non-voisés :** cette mesure utilisée dans [DOI et collab., 2014; TANAKA et collab., 2014], permet d'évaluer la précision de la prédiction des caractéristiques d'excitation. Les auteurs de ces études mentionnent avoir utilisé les coefficients de corrélation et le taux d'erreur des segments voisés/non-voisés sur la F0 et les composantes apériodiques (l'enveloppe spectrale supérieure [OHTANI et collab., 2006]) entre la voix convertie et voix cible. Les coefficients de corrélation permettent de donner une mesure synthétique de l'intensité de la relation entre ces deux paramètres.

- **Mesure de log-vraisemblance :** est une mesure estimée à l'aide d'un système de reconnaissance du locuteur, permettant d'évaluer le degré de rapprochement (identification) du locuteur source/cible de la voix transformée. Cette mesure est définie par l'équation suivante :

$$\mathrm{V_L(Y)} = \log(p(\mathrm{Y/M}_c)) - \log(p(\mathrm{Y/M}_s)) \tag{4.15}$$

  Avec $p(\mathrm{Y/M}_s)$ et $p(\mathrm{Y/M}_c)$ représentant les probabilités que le signal converti y ait été prononcé respectivement par le locuteur source ou cible, M_s est le modèle de la voix source et M_c le modèle de la voix cible.

4.4.2 Évaluation subjective

L'évaluation subjective doit faire appel à au moins deux types de données qui sont les données converties et les données issues du locuteur cible. Les tests subjectifs (d'écoute) les plus utilisés sont :

- **Test ABX :** c'est un test qui permet d'évaluer l'identité de la voix. Trois voix sont présentées aux auditeurs, la voix A, B et X respectivement du locuteur source, cible

et de la voix convertie. Ces auditeurs jugent par une note le degré de rapprochement de la voix convertie X aux deux autres voix des locuteurs A et B. Cette note est binaire est peut être graduellement étendue à 5 niveaux (voir tableau 4.1).

Note	1	2	3	4	5
Jugements	X est le locuteur A	X est similaire au locuteur A	X n'est ni A ni B	X est similaire au locuteur B	X est le locuteur B

TABLEAU 4.1: *Note graduelle à 5 niveaux concernant le test ABX*

Différents travaux ont utilisé le test ABX comme [ABE et collab., 1988; KAIN et MACON, 1998; STYLIANOU et collab., 1998]. Il faut noter que le test ABX est inadéquat dans le cas d'une conversion de voix inter-genre, c'est-à-dire, la conversion homme/femme ou femme/homme.

- **Test MOS (Mean Opignion Score) :** c'est un test qui permet d'évaluer la qualité de la voix convertie resynthétisée. Les auditeurs jugent par une note la qualité de la parole convertie sur une échelle numérique. Cette échelle va de un, pour la plus mauvaise qualité, jusqu'à cinq pour une qualité excellente ((2) médiocre (3) moyenne et (4) bonne qualité)). Le score moyen est utilisé pour décider de la qualité de la parole convertie. Ce test a été utilisé dans plusieurs travaux de recherche, comme par exemple [KAIN et MACON, 1998] et [TODA, 2003].

4.5 Notre système hybride pour l'amélioration de la reconnaissance de la parole œsophagienne

Nous décrirons dans cette section, la théorie et la mise en œuvre de notre système hybride [LACHHAB et collab., 2015], proposé pour l'amélioration de la parole œsophagienne. Ce système hybride basé sur la conversion de voix par des GMMs, vise à compenser l'information déformée présente dans les vecteurs acoustiques de la parole œsophagienne. La parole œsophagienne "source" est convertie en parole laryngée "cible" en utilisant une fonction de conversion estimée statistiquement à l'aide d'un algorithme itératif simple et rapide. Contrairement aux recherches antérieures, nous n'avons pas appliqué un algorithme de re-synthèse vocale pour reconstruire le signal de la parole convertie. Les vec-

teurs Mel cepstraux convertis sont utilisés directement comme entrée dans notre système de reconnaissance automatique de la parole œsophagienne (voir section 3.4) pour évaluer l'amélioration de l'extraction phonétique après conversion. En outre les vecteurs acoustiques MFCC sont linéairement transformés par la méthode HLDA (voir section 2.8.2) pour réduire leur dimension dans un espace restreint ayant de bonnes propriétés discriminantes. Les résultats expérimentaux démontrent que notre système hybride proposé fournit une amélioration absolue du taux de reconnaissance phonétique (Accuracy) de 3.40% par rapport au système de référence qui fonctionne sans transformation HLDA ni conversion de voix. La figure 4.6, illustre le schéma fonctionnel de notre système hybride de correction proposé.

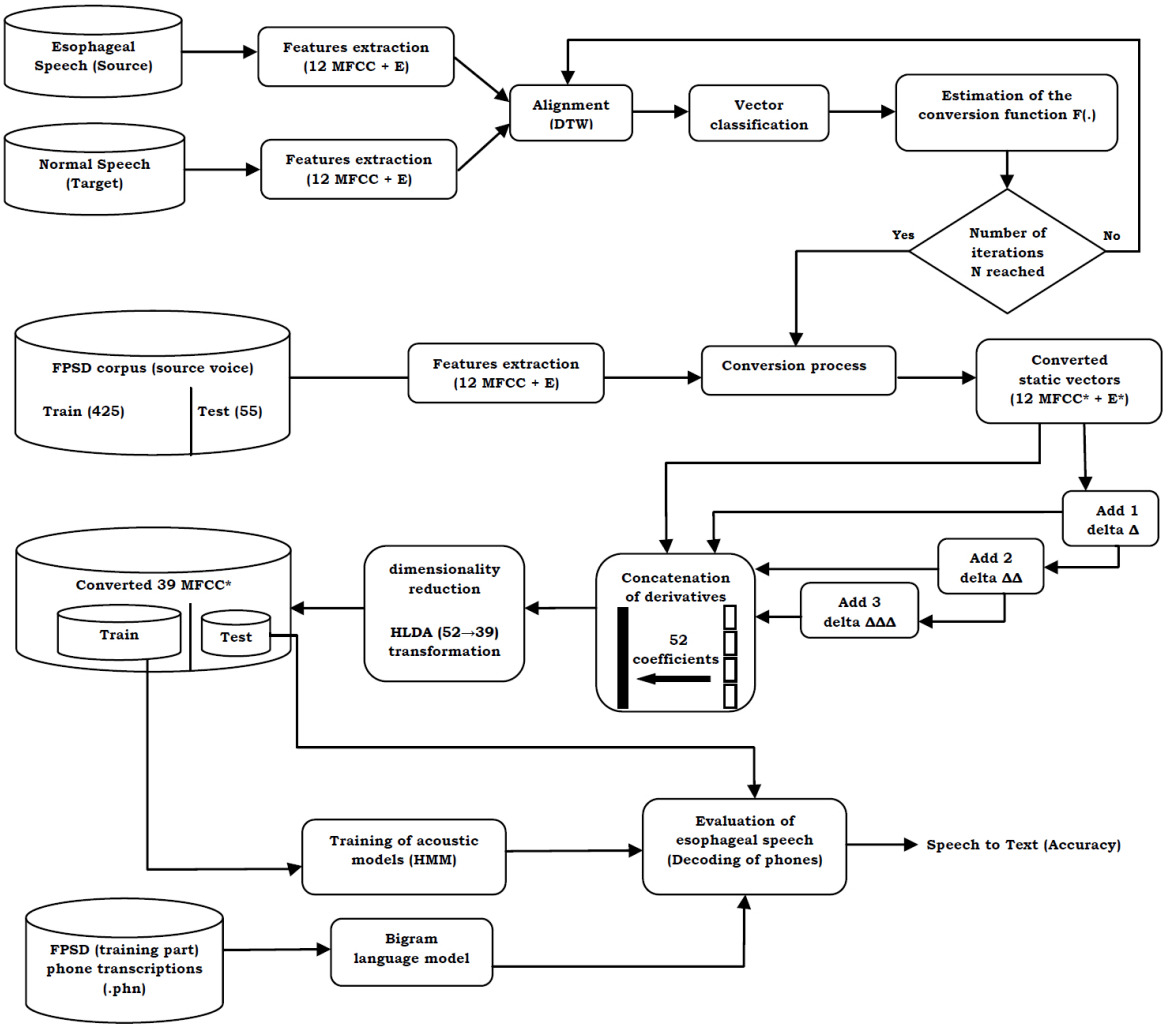

FIGURE 4.6: *Le schéma fonctionnel du système hybride proposé pour améliorer la reconnaissance de la parole œsophagienne.*

4.5.1 Extraction des vecteurs acoustiques

On dispose de deux corpus parallèles de voix source (œsophagienne) et cible (laryngée) dont les phrases enregistrées possèdent le même contenu phonétique. Cependant, chaque locuteur se caractérise par son style d'élocution : rythme, timbre et débit d'élocution. Ainsi la taille ou le nombre d'échantillons des phrases des deux corpus sont forcement différents même dans le cas où les deux locuteurs essayent de s'imiter l'un l'autre. Pour remédier à cette situation, nous avons normalisé dans une étape de pré-traitement les fichiers audio des phrases (cible) afin qu'elles aient les mêmes nombres d'échantillons que leurs correspondantes phrases (source). Ce pré-traitement a été effectué par le logiciel open source "SoX" (en anglais Sound eXchange), qui est un outil de manipulation et de traitement des fichiers sonores. En général, la mise en œuvre d'un système de conversion de voix n'entraîne pas l'application d'une normalisation en nombre d'échantillons sur les fichiers sonores. Toutefois, sa mise en œuvre permet d'améliorer l'alignement DTW des vecteurs source→cible. Ensuite, ces signaux de la parole issus des locuteurs source et cible (normalisés) subissent une phase de paramétrisation. Le but de cette paramétrisation est d'extraire les vecteurs cepstraux MFCC. Dans ce traitement, le signal de parole est échantillonné à 16 kHz avec une préaccentuation de 0.97. Une fenêtre de Hamming de 25 ms décalée toutes les 10 ms est utilisée pour obtenir des sections de courte durée à partir desquelles les coefficients cepstraux sont extraits. Les 12 premiers coefficients cepstraux (c_1 à c_{12}) sont concaténés avec le logarithme de l'énergie de la trame pour former des vecteurs MFCC statique de 13 coefficients (12MFCC + E). Ces coefficients sont calculés en utilisant une fenêtre de Hamming de 25 ms décalée toutes les 10 ms et à l'aide d'un banc de 26 filtres dans une échelle de fréquence Mel.

Les coefficients différentiels d'ordre 1,2 et 3 ($\Delta, \Delta\Delta$ et $\Delta\Delta\Delta$) ne sont pas utilisés dans le processus de conversion. Ils sont calculés directement à partir des coefficients statiques des vecteurs MFCC convertis, pour servir d'entrées au système RAP. Cette procédure est importante dans le but de conserver les informations dynamiques des dérivées qui peuvent être perdues lors de la conversion.

4.5.2 L'alignement DTW

Le principe de l'alignement DTW consiste à mettre en correspondance les deux séquences de vecteurs X_N et Y_N (source et cible). Cependant l'inconvénient de l'algorithme DTW, dans sa version classique est qu'il nécessite un temps de calcul important qui augmente en fonction du nombre de vecteurs N traités. Pour cette raison, nous avons implémenté une variante de cet l'algorithme DTW, en réduisant l'ensemble des alignements possibles dans la recherche du chemin optimal. La région de contrainte dans laquelle peuvent apparaître les couples alignés est similaire au parallélogramme d'Itakura [ITAKURA, 1975] (voir la figure 4.7). Cette variante consiste à diminuer la complexité en limitant l'espace de recherche autour de la diagonale.

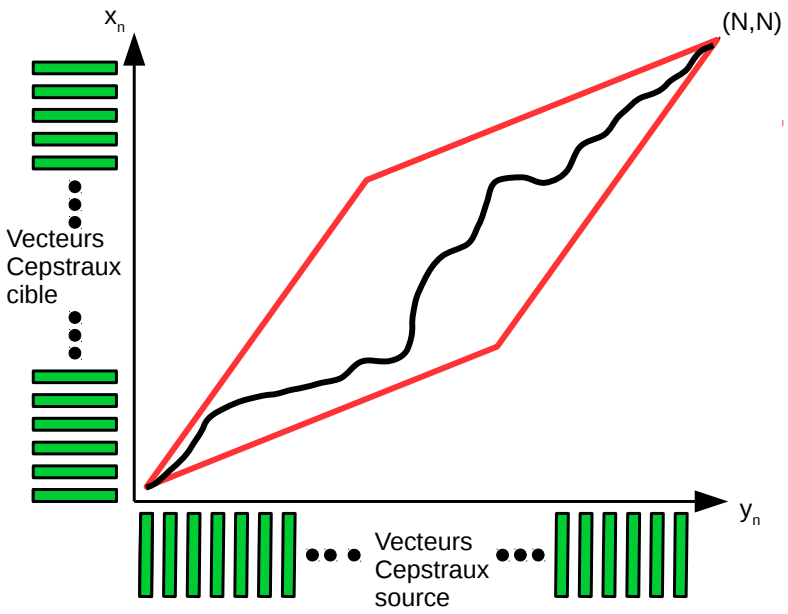

FIGURE 4.7: *Le parallelogramme utilisé dans l'alignement temporel par la DTW.*

Trois chemins sont possibles :

- Le chemin 1 passe par les couples de vecteurs $(i-2, j-1)$ et $(i-1, j)$.
- Le chemin 2 passe par les couples de vecteurs $(i-1, j-1)$.
- Le chemin 3 passe par les couples de vecteurs $(i-1, j-2)$ et $(i, j-1)$.

En sortie de cet algorithme d'alignement optimal, nous obtenons une séquence de couples correspondant à un ensemble de vecteurs cepstraux source et cible appariés. Cette liste de couples de correspondance sera utilisée dans la phase d'apprentissage pour calculer les paramètres de la GMM et ainsi estimer la fonction de conversion.

4.5.3 Apprentissage de la fonction de conversion

Nous avons utilisé le modèle GMM décrit par [STYLIANOU et collab., 1998] et amélioré par [KAIN et MACON, 1998] puis par [WERGHI et collab., 2010]. La procédure d'estimation des paramètres GMM pour le calcul de la fonction de conversion est basée sur l'algorithme itératif ISE2D (Iterative Statistical Estimation Directly from Data) proposé par Wergui.

Nous supposons avoir deux séquences de vecteurs cepstraux MFCC avec un nombre identique (normalisation), $X_N^d = [x_1^d, x_2^d, \ldots, x_n^d]$ et $Y_N^d = [y_1^d, y_2^d, \ldots, y_n^d]$ source et cible, N étant le nombre de vecteurs et d étant leur dimension. Ces deux séquences possédant le même contenu phonétique ont été alignés temporellement par l'algorithme DTW décrit ci-dessus, pour associer les vecteurs source à leurs vecteurs cible correspondants. Les vecteurs appariés (source/cible) sont ensuite concaténés conjointement dans un seul vecteur étendu $Z_N^d = [X_N^d Y_N^d]$ avant la classification. Cette concaténation est utilisée pour modéliser conjointement un GMM, qui dépend à la fois des paramètres source et cible (α_i,μ_i^x, μ_i^y, Σ_i^{xx}, Σ_i^{yx}). Ces paramètres sont calculés grâce à une classification vectorielle. La classification vectorielle est une étape nécessaire afin de diviser l'espace des vecteurs MFCC en classes ou régions. Chaque classe est caractérisée par un centroïde (vecteur moyen). L'algorithme K-moyens (en anglais K-means) [MACQUEEN et collab., 1967], a été utilisé pour effectuer cette classification vectorielle. Son choix a été guidé par sa simplicité et sa rapidité d'excecution et par le fait qu'il autorise la détermination d'un nombre quelconque de classes contrairement à l'algorithme LBG (Linde, Buzo et Gray) [LINDE et collab., 1980].

Les paramètres GMM sont estimés itérativement par l'algorithme ISE2D. Cet algorithme est moins coûteux en temps de calcul et donne de meilleur résultats, contrairement à l'algorithme standard EM. [WERGHI et collab., 2010] ont montré que EM pouvait être avantageusement remplacé par l'algorithme itératif ISE2D. D'après le schéma fonctionnel, l'algorithme ISE2D incorpore l'alignement DTW et la classification vectorielle k-

means de l'espace des vecteurs d'apprentissage Z_n pour estimer statistiquement les paramètres GMM comme suit :

- Le poids α_i de la distribution normale est estimé comme étant le rapport entre $N_{s,i}$ le nombre de vecteurs source (s) de la classe i, et N_s qui représente le nombre total de vecteurs source :

$$\alpha_i = \frac{N_{s,i}}{N_s} \tag{4.16}$$

- Le vecteur moyen (centroïde) μ_i^x source et vecteur moyen μ_i^y cible de la classe i sont calculés comme suite :

$$\mu_i^x = \frac{\sum_{n=1}^{N_{s,i}} x_n}{N_{s,i}} \tag{4.17}$$

Et

$$\mu_i^y = \frac{\sum_{n=1}^{N_{c,i}} y_n}{N_{c,i}} \tag{4.18}$$

Où x_n,y_n et $N_{c,i}$ représentent le $n^{\text{ème}}$ vecteur source, le $n^{\text{ème}}$ vecteur cible et le nombre de vecteurs de la classe i.

- Finalement, la matrice de covariance Σ_i^{xx} et la matrice de covariance croisée Σ_i^{yx} des vecteurs cible/source de la classe i sont calculées par la formule classique :

$$\Sigma_i^{xx} = \mathbb{E}[(x_i - \mu_i^x)((x_i - \mu_i^x)')] \tag{4.19}$$

Et

$$\Sigma_i^{yx} = \mathbb{E}[(y_i - \mu_i^y)((x_i - \mu_i^x)')] \tag{4.20}$$

Cet algorithme propose dans la première itération, d'appliquer l'alignement DTW entre les vecteurs source X_N et cible Y_N. A partir de la seconde itération, l'alignement est réalisé entre les vecteurs convertis $\hat{Y}_N$ et les vecteurs cibles Y_N dans le but d'affiner le chemin d'alignement temporel.

Une fois les paramètres des GMMs calculés, la fonction de conversion précédemment définie par la formule 4.4 est appliquée au $n^{\text{ème}}$ vecteur MFCC source x_n afin de prédire le $n^{\text{ème}}$ vecteur converti $\hat{y}_n$. Ce processus de conversion est défini par l'équation suivante :

$$\hat{y}_n = \mathscr{F}(x_n) \tag{4.21}$$

Algorithme 4.1 : *K-means*

---

1. Initialisation :

À l'instant $t=0$, choix aléatoire (ou guidé) de p centroides.
Une distorsion initiale $\mathrm{D}^{(0)}=\infty$ et un seuil $\varepsilon>0$.

2. Calcul des distances entre les vecteurs d'apprentissage $\mathrm{Z_N}$ et les centroides μ_p.

3. Mise à jour des classes C_i (association des vecteurs au centroides les plus proches), avec $z_i \in \mathrm{C}_i$ si $\mathrm{D}ist(z_i,\mu_i) < \mathrm{D}ist(z_i,\mu_l), \forall l \in [1,p], l \neq i$.

4. Remplacer chaque centroide μ_i par le centre de gravité de la classe C_i.

5. Conditions d'arrêt :

Calcul de la distorsion moyenne D_m pour la partition obtenue, avec $\mathrm{D}_m^{(t)} = \frac{1}{\mathrm{N}} \sum_{n=1}^{\mathrm{N}} [\min_{i=1}^{p} \mathrm{D}ist(z_n,\mu_i)]$.
Si $\dfrac{(\mathrm{D}_m^{(t-1)} - \mathrm{D}_m^{(t)})}{\mathrm{D}_m^{(t)}} < \varepsilon$
Alors l'algorithme est terminé (pas de changement au niveau des classes)
Sinon $t=t+1$ et aller à l'étape **2.**

Algorithme 4.2 : *LBG*

1. Initialisation :

Le centroide initial μ_0 $(p=1)$ de l'ensemble des vecteurs d'apprentissage Z_N est calculé à l'instant $t=0$.

2. Eclatement "Splitting" des centroides.

$t = t+1$

Pour $i=1$ **à** p **faire :**

$$\mu_{2i-1}(t) = \mu_{i-1}(t-1) + V$$
$$\mu_{2i}(t) = \mu_{i-1}(t-1) - V$$

avec V un vecteur aléatoire de variance adaptée aux vecteurs associés à μ_i. Multiplier p par 2.

3. Faire tourner les K-means sur T itérations.

4. Tant que p n'a pas atteint la valeur souhaitée, aller à l'étape **2.**

Notre système hybride de correction a été proposé pour améliorer le décodage de la parole œsophagienne. Ce système de correction combine deux approches différentes (hybride) : la conversion statistique de la voix qui transforme la parole œsophagienne source en parole laryngée cible, avec un système de reconnaissance automatique de la parole, basé sur l'approche statistique HMM/GMM. Notre système hybride ne nécessite pas l'application d'un algorithme de re-synthèse vocale pour reconstruire la parole convertie afin de juger ou évaluer sa qualité et son intelligibilité. Notre objectif principal est d'améliorer la reconnaissance automatique de phonèmes de cette parole œsophagienne. La parole convertie n'est pas plus intelligible que la parole originale (œsophagienne) mais permet de réaliser une meilleure reconnaissance (Speech-to-Text). La principale contribution de notre approche est la conversion des vecteurs cepstraux MFCC (source/cible) qui sont directement utilisés en entrée du système de reconnaissance de la parole œsophagienne décrit dans la section 3.4. Cette méthode instrumentale, rapide et peu coûteuse en ressources humaines, nous a permis d'améliorer la reconnaissance de cette parole patholo-

gique. La transformation HLDA appliquée aux MFCC* [1] a permis d'améliorer les performances du système.

4.6 Expériences et résultats

Afin de convertir la parole œsophagienne en "parole normale" nous avons enregistré 50 phrases œsophagiennes et laryngées respectivement prononcées par une personne laryngectomisée masculin français (le même qui a participé à la création de notre base de données FPSD) et un locuteur masculin français ayant une voix non-pathologique (laryngée). Ces nouveaux enregistrements n'appartiennent pas au corpus FPSD. Ils ont été enregistrés dans le but d'estimer statistiquement la fonction de conversion. Au cours de la première itération de l'apprentissage, l'alignement DTW est appliqué sur les vecteurs source X_N et cible Y_N contenant les 13 coefficients statiques. A partir de la deuxième itération, l'alignement DTW est réalisé entre les vecteurs statiques convertis $\hat{Y}_N$ et les vecteurs cible Y_N dans le but d'affiner la liste de correspondance (mapping). La fonction de conversion est estimée en utilisant 64 classes. Nous avons effectué trois expériences à l'aide du système de reconnaissance de phonèmes de la parole œsophagienne. L'objectif de ces expériences est de mesurer le degré d'amélioration obtenu par notre système hybride (l'expérience de conversion précédemment décrite ne change pas).

Dans la première expérience, la même formule de régression HTK décrite dans la section 1.4.2 a été utilisée pour calculer les dérivées d'ordre 1 et 2 à partir des vecteurs statiques convertis. Le but de cette expérience est d'ajouter les informations dynamiques et avoir de nouveaux vecteurs de dimension = 39 (12 MFCC*, E*; 12 ΔMFCC*, ΔE*; 12 $\Delta\Delta$MFCC*, $\Delta\Delta$E*) représentant la dimensionnalité de référence).

Dans la deuxième expérience, une autre dérivée ($\Delta\Delta\Delta$) est ajoutée et concaténée dans l'espace des vecteurs afin d'augmenter leurs nombre de coefficients à d = 52 (12 MFCC*, E*; 12 ΔMFCC*, ΔE*; 12 $\Delta\Delta$MFCC*, $\Delta\Delta$E*; 12 $\Delta\Delta\Delta$MFCC*, $\Delta\Delta\Delta$E*).

Dans la troisième expérience, l'espace de 52 coefficients utilisés dans l'expérience 2 est réduite à 39 coefficients en utilisant la transformation HLDA (52 → 39) en vue d'améliorer l'information discriminante et de réduire la dimensionnalité de l'espace.

1. MFCC* : Signifie vecteurs MFCC convertis

Les taux de reconnaissance de phonème (Accuracy) et les taux corrects, sont calculés à l'aide de notre système de reconnaissance de la parole œsophagienne (voir la section 3.4) dans le but d'évaluer la conversion des vecteurs MFCC.

Le tableau 4.2 présente les résultats des trois expériences décrites ci-dessus sur les vecteurs MFCC* de la partie de test de notre propre base de données FPSD contenant 55 phrases.

36 HMMs monophone avec 16 Gaussiennes par état + Bigramme	**Accuracy (%)**	**Correct (%)**
Expérience 1 : 39 coefficients MFCC∗	63.48	68.58
Expérience 2 : 52 coefficients MFCC∗	61.78	67.36
Expérience 3 : 39 coefficients HLDA (52 → 39)	**65.29**	**69.85**

TABLEAU 4.2: *L'apport des coefficients différentiels et de la transformation HLDA sur le taux de reconnaissance phonétique (Accuracy) obtenu en utilisant les vecteurs* MFCC* *convertis de la partie Test de notre base de données FPSD*

Les résultats exposés dans le tableau 3.2 de la section 3.4.4, présentent les taux de reconnaissance de phonèmes pour les trois expériences décrites ci-dessus, sur la partie test de notre corpus originale FPSD. On peut observer à partir des résultats de l'expérience 3 (tableau 4.2), que le système hybride proposé fournit une amélioration du taux de reconnaissance de phonèmes par une augmentation absolue de 3.40%. Le fait que les performances de notre système après conversion aient été améliorées valide le caractère hybride du logiciel proposé.

Ainsi nous avons démontré que la transformation HLDA et la technique de conversion de la voix peuvent conjointement améliorer les propriétés discriminantes des trames cepstrales calculées.

4.7 Conclusion

Nous avons décrit dans ce chapitre les étapes de construction de notre système hybride de correction, capable d'améliorer la reconnaissance de la parole œsophagienne. Ce système hybride est basé sur une conversion statistique GMM simplifiée, qui projette les vecteurs de la parole œsophagienne dans un espace moins "perturbé" relatif à la pa-

role laryngée. Nous n'utilisons pas un algorithme de re-synthèse vocale pour reconstruire le signal sonore de la parole convertie, parce que notre système de reconnaissance de phonèmes utilise directement les vecteurs Mel cepstraux convertis comme entrées. Nous avons aussi projeté ces vecteurs MFCC* convertis par la transformation HLDA dans un espace restreint ayant de bonnes propriétés discriminantes. Les taux de décodage obtenus, démontrent que le système hybride proposé permet une amélioration significative de la reconnaissance automatique de la parole œsophagienne. Nous envisageons dans nos futurs travaux, de réaliser un dispositif portable qui effectuera la reconnaissance de la parole œsophagienne ainsi que la reconstruction du signal de la parole reconnue en utilisant un synthétiseur texte-parole (Text-to-Speech). Un tel dispositif permettrait aux personnes laryngectomisées une communication orale plus facile avec d'autres personnes. Néanmoins, le système de reconnaissance de la parole œsophagienne devrait être en mesure de restaurer une quantité conséquente d'information phonétique (Speech-to-Text). Pour cette raison, nous avons l'intention d'étendre notre corpus FPSD afin de rendre possible l'utilisation des modèles HMM dépendant du contexte (triphones). De plus, nous envisageons de remplacer notre méthode de conversion de la voix par un algorithme similaire à celui de Toda [TODA et collab., 2007] afin d'améliorer le processus de conversion de la voix et conséquemment la précision de la reconnaissance de la parole.

Conclusion générale et perspectives

Conclusion générale

L'objectif de cette thèse est la réalisation d'un système de reconnaissance automatique de la parole œsophagienne (alaryngée). L'étude de ce type de parole pose plusieurs problèmes difficiles : 1) Les corpus de la parole œsophagienne existants ne sont pas dédiés à la reconnaissance, à cause d'un manque de données (uniquement quelques dizaines de phrases enregistrées pour des besoins ponctuels d'une étude) ; 2) Contrairement à la parole laryngée (normale), la parole œsophagienne (alaryngée) est caractérisée par un bruit spécifique élevé, une faible intelligibilité et une fréquence fondamentale instable. Toutes ces caractéristiques permettent de produire une voix rauque, grinçante et non naturelle, difficile à comprendre ; 3) les systèmes de reconnaissance automatique de la parole laryngée peuvent être adaptés à cette parole alaryngée mais avec des pertes en performance ; 4) La difficulté de compenser les distorsions spectrales ou cepstrales entre ces deux types de parole ; 5) L'extraction des paramètres de voisement pour la re-synthèse de la parole comporte certaines déficiences. Pour apporter une solution à tous ces défis, nous avons dirigé cette thèse selon plusieurs axes :

Le premier, concerne l'étude et l'implémentation d'un système de reconnaissance automatique de la parole laryngée en utilisant les modèles de Markov cachées. Dans ce sens, trois systèmes de reconnaissance de la parole continue ont été créé. Le premier nommé "SPIRIT" utilise une méthode simple d'apprentissage basée sur l'estimation directe des paramètres à partir des données en utilisant les algorithmes LBG et Viterbi au lieu de la procédure classique de Baum-Welch. Dans ce système, nous avons proposé un modèle de durée d'émission des observations pour les modèles phonétiques indépendants du contexte. Ce modèle de durée est basée sur une distribution normale capable d'améliorer

le taux de reconnaissance de ce système. Les deux autres systèmes créés sont plus performants. Ils ont été implémentés à l'aide de la plate-forme HTK, l'un est basé sur des modèles phonétiques monophones et l'autre plus robuste car il tient compte du contexte phonétique gauche et droit (triphones).

Le deuxième axe suivi dans cette thèse est lié à la conception de notre propre base de données de la parole œsophagienne. Cette base de données que nous avons nommé FPSD contient 480 phrases prononcées par un locuteur laryngectomisé qui a acquis la voix œsophagienne après une rééducation vocale. Ces 480 phrases ont été segmentées manuellement en mots et en phonèmes afin de faciliter l'apprentissage et le décodage du système de reconnaissance.

Le troisième axe est relatif à l'adaptation et l'application du système de reconnaissance de la parole laryngée à la parole œsophagienne en utilisant cette base de données (FPSD). Le système le plus à même pour accomplir cette tâche est le système de reconnaissance monophones (HTK), car notre corpus ne contient pas assez de données pour faire l'apprentissage des modèles phonétiques triphones. La transformation discriminante HLDA a été appliquée sur les vecteurs acoustiques pour améliorer l'information discriminante entre les classes phonétiques et afin d'améliorer le décodage de la parole œsophagienne.

Le dernier axe poursuivi dans cette thèse réside dans la réalisation d'un système hybride (correction = conversion + reconnaissance) capable de corriger les distorsions présentes dans le signal de la parole œsophagienne. Ce système hybride de correction, basé sur la conversion de la voix œsophagienne⟶laryngée, a pour objectif d'améliorer la reconnaissance de cette parole œsophagienne.

Perspectives

Le travail présenté dans ce manuscrit est une démarche pour répondre à la problématique que nous nous somme fixée. Les solutions proposées sont certainement incomplètes mais laissent entrevoir de nombreuses perspectives. Il va falloir, dans un premier temps, étendre notre corpus FPSD afin de rendre possible l'utilisation des modèles pho-

nétiques dépendants du contexte à partir de notre système de reconnaissance triphones. Cette approche permettra sans aucun doute d'améliorer le taux de reconnaissance de phonèmes d'environ 5 à 7%.

Dans un deuxième temps, notre méthode simple de conversion de la voix utilisée dans le système hybride de correction de la parole œsophagienne, peut être remplacé par d'autres techniques plus sophistiquées, comme par exemple l'algorithme de conversion de la voix de Toda [TODA et collab., 2007] ou l'approche EigenVoice proposée dans [TODA et collab., 2006].

Nous envisageons aussi la possibilité d'utiliser un synthétiseur texte-parole performant (Text-to-Speech), afin de reconstruire une parole laryngée à partir de l'information phonétique ou lexicale extraite grâce au décodage de notre système de reconnaissance. Ce processus complet permettrait sans aucun doute aux personnes laryngectomisées, une communication orale plus facile avec d'autres personnes.

Publications de l'auteur

Journaux Internationaux

Othman LACHHAB, Joseph Di MARTINO, El Hassane Ibn ELHAJ et Ahmed HAMMOUCH, "A preliminary study on improving the recognition of esophageal speech using a hybrid system based on statistical voice conversion", SpringerPlus journal, vol. 4, n° 1, doi :10.1186/s40064-015-1428-2, p. 1–14, ISSN 2193-1801, October 2015.

Conférences Internationales avec comité de lecture

Othman LACHHAB, Joseph Di MARTINO, El Hassane Ibn ELHAJ et Ahmed HAMMOUCH, "Improving the recognition of pathological voice using the discriminant HLDA transformation", In third IEEE International Colloquium in Information Science and Technology (CIST), p. 370–373, October 2014, Tetuan, Morocco.

Othman LACHHAB, Joseph Di MARTINO, El Hassane Ibn ELHAJ et Ahmed HAMMOUCH, "Real time context- independent phone recognition using a simplified statistical training algorithm", 3rd International Conference on Multimedia Computing and Systems - ICMCS'12. URL https ://hal.inria.fr/hal-00761816/document, May 2012, Tanger, Morocco.

Othman LACHHAB, El Hassane Ibn ELHAJ, "Improved feature vectors using N-to-1 Gaussian MFCC transformation for automatic speech recognition system", In the 5th International Conference on Multimedia Computing and Systems (ICMCS'16) – IEEE Conference, p. 76-81, 29 September 2016, Marrakech, Morocco.

Journées nationales

Othman LACHHAB, Joseph Di MARTINO, El Hassane Ibn ELHAJ et Ahmed HAMMOUCH, "Reconnaissance de la parole continue indépendant du locuteur en utilisant des CI-CDHMMs", Séminaire Oesovox à l'INPT, 2011, Rabat, MAROC.

Bibliographie

ABE, M., S. NAKAMURA, K. SHIKANO et H. KUWABARA. 1988, «Voice conversion through vector quantization», *In Proc. ICASSP*, p. 655–658. 84, 86, 94, 95

ALI, R. H. et S. B. JEBARA. 2006, «Esophageal speech enhancement using excitation source synthesis and formant patterns modification», *In Proc. Int. Conf. on Signal-Image Technology & Internet Based Systems (SITIS)*, p. 315–324. 80

BAHL, L., P. BROWN, P. V. DE SOUZA et R. MERCER. 1986, «Maximum mutual information estimation of hidden markov model parameters for speech recognition», dans *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP 86*, vol. 11, p. 49–52.
doi:10.1109/ICASSP.1986.1169179. 20

BAHL, L., P. BROWN, P. V. DE SOUZA et R. MERCER. 1989, «A tree-based statistical language model for natural language speech recognition», *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, nº 7, p. 1001–1008. 25

BAKER, J. 1975, «The dragon system–an overview», *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 23, nº 1, p. 24–29.
doi:10.1109/TASSP.1975.1162650. 17

BAUM, L. E. 1972, «An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes», *Inequalities*, vol. 3, p. 1–8. 20, 21, 76

BELLANDESE, M. H., J. W. LERMAN et H. R. GILBERT. 2001, «An acoustic analysis of excellent female esophageal, tracheoesophageal, and laryngeal speakers», *Journal of Speech, Language and Hearing Research*, vol. 44, nº 1, p. 1315–1320. 66

BOLL, S. F. 1979, «Suppression of acoustic noise in speech using spectral subtraction», *Acoustics, Speech and Signal Processing, IEEE Transactions*, vol. 27, nº 2, p. 113–120. 82

BURGET, L. 2004, «Combination of speech features using smoothed heteroscedastic linear discriminant analysis», *In 8th International Conference on Spoken Language Processing*, p. 2549–2552. 57

BÉCHET, F. 2001, «LIA–PHON : Un système complet de phonétisation de textes», *Revue Traitement Automatique des Langues (TAL)*, p. 47–67. 24

CAROL, Y., V. CHARI, J. MACAUSLAN, C. HUANG et M. WALSH. 1998, «Enhancement of electrolaryngeal speech by adaptive filtering», *Journal of Speech, Language and Hearing Research,*, vol. 41, nº 1, p. 1253–1264. 66

CHOMSKY, N. 1965, «Aspects of the theory of syntax», *MIT Press, Cambridge*. 25

COLE, D., S. SRIDHARAN et M. GEVA. 1997, «Application of noise reduction techniques for alaryngeal speech enhancement», *Speech & Image Process. for Computing & Telecommun.*, p. 491–494. 81

DAVIS, S. et P. MERMELSTEIN. 1980, «Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences», *In IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, nº 4, p. 357–366. doi:10.1109/TASSP.1980.1163420. 10, 14, 75

DEMPSTER, A., N. LAIRD et D. RUBIN. 1977, «Maximum likelihood from incomplete data via the em algorithm», *Journal of the Royal Statistical Society. Series B (Methodological*, vol. 39, nº 1, p. 1–38. 89

DESAI, S., A. W. BLACK, B. YEGNANARAYANA et K. PRAHALLAD. 2010, «Spectral mapping using artificial neural networks for voice conversion», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, nº 5, p. 954–964. doi:10.1109/TASL.2010.2047683. 88, 93

DIBAZAR, A., T. W. BERGER et S. NARAYANAN. 2006, «Pathological voice assessment», *Engineering in Medicine and Biology Society. EMBS 06. 28th Annual International Conference of the IEEE*, p. 1669–1673. 60, 74

DOI, D., T. TODA, K. NAKAMURA, H. SARUWATARI et K. SHIKANO. 2014, «Alaryngeal speech enhancement based on one-to-many eigenvoice conversion», *IEEE Trans. Audio. Speech Language*, vol. 22, nº 1, p. 172–183. 6, 82, 83, 90, 93, 94

EN-NAJJARY, T. 2005, *Conversion de voix pour la synthèse de la parole*, thèse de doctorat, Traitement du signal et de l'image. Université Rennes 1. 83

FU, K. 1971, «On syntactic pattern recognition and stochastic languages», *inProc. International Conference on Frontiers of Pattern Recognition, Hawaii.* 25

FURUI, S. 1986, «Speaker-independent isolated word recognition using dynamic features of speech spectrum», *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 34, nº 1, p. 52–59.
doi:10.1109/TASSP.1986.1164788. 16

GALES, M. J. F. 1999, «Semi-tied covariance matrices for hidden markov models», *IEEE Transactions on Speech and Audio Processing*, vol. 7, nº 3, p. 272–281. 57

GARCIA, B., J. VICENTE et E. ARAMENDI. 2002, «Time-spectral technique for esophageal speech regeneration», *Biosignal Analysis of biomedical signals and images*, p. 113–116. 81

GARCIA, B., J. VICENTE, I. RUIZ, A. ALONSO et E. LOYO. 2005, «Esophageal voices : Glottal flow restoration», *In Proc. ICASSP*, p. 141–144. 81

GAROFOLO, J. S., L. F. LAMEL, W. M. FISHER, J. G. FISCUS, D. PALLETT et N. L. DAHLGREN. 1993, *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM*. NTIS order number PB91-100354. 6, 17, 31, 68

GAUVAIN, J. et C.-H. LEE. 1994, «Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains», *Speech and Audio Processing, IEEE Transactions on*, vol. 2, nº 2, p. 291–298.
doi:10.1109/89.279278. 20

HAEB-UMBACH, R. et H. NEY. 1998, «Linear discriminant analysis for improved large vocabulary continuous speech recognition», *In Proc. ICASSP*, p. 13–16. 55, 56

HARRIS, F. 1978, «On the use of windows for harmonic analysis with the discrete fourier transform», *Proceedings of the IEEE*, vol. 66, nº 1, p. 51–83.
doi:10.1109/PROC.1978.10837. 16

HERMANSKY, H. 1990, «Perceptual linear predictive (PLP) analysis for speech», *journal of acoustical society of america*, vol. 87, p. 1738–1752.
doi:10.1121/1.399423. 15

HISADA, A. et H. SAWADA. 2002, «Real-time clarification of esophageal speech using a comb filter», *International Conference on Disability, Virtual Reality and Associated Technologies*, p. 39–46. 81

ITAKURA, F. 1975, «Minimum prediction residual principle applied to speech recognition», *Speech communication journal*, vol. 23, nº 1, p. 67–72. 98

JELINEK, F. 1976, «Continuous speech recognition by statistical methods», *Proceedings of the IEEE*, vol. 64, nº 4, p. 532–556.
doi:10.1109/PROC.1976.10159. 13, 17, 25

JELINEK, F. et R. L. MERCER. 1980, «Interpolated estimation of markov source parameters from sparse data», *Proc. Workshop Pattern Recognition in Practice*, p. 381–397. 25

JELINEK, F., R. L. MERCER, L. R. BAHL et J. K. BAKER. 1977, «Perplexity a measure of the difficulty of speech recognition tasks», *journal of acoustical society of america*, vol. 62, p. S63.
doi:10.1121/1.2016299. 26

JOUVET, D., L. MAUUARY et J. MONNÉ. 1991, «Automatic adjustments of the structure of markov models for speech recognition applications», *proceeding EUROSPEECH 91*, p. 927–930. 43

JUANG, B. et L. RABINER. 1985, «Mixture autoregressive hidden markov models for speech signals», *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, nº 6, p. 1404–1413.
doi:10.1109/TASSP.1985.1164727. 19

KAIN, A. et M. MACON. 1998, «Spectral voice conversion for text-to-speech synthesis», *In Proc. ICASSP*, p. 285–288. 83, 84, 86, 89, 95, 99

KATZ, S. 1987, «Estimation of probabilities from sparse data for the language model component of a speech recognizer», *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, nº 3, p. 400–401. 25

KAWAHARA, H. 1997, «Speech representation and transformation using adaptive interpolation of weighted spectrum : vocoder revisited», *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 2, p. 1303–1306. doi:10.1109/ICASSP.1997.596185. 92

KAWAHARA, H., I. MASUDA-KATSUSE et A. DE CHEVEIGNE. 1999, «Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction : Possible role of a repetitive structure in sounds», *Speech communication journal*, vol. 27, nº 3, p. 187–207. 92

KUHN, R. et R. D. MORI. 1990, «A cache-based natural language model for speech recognition», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, nº 6, p. 570–583. 25

KUMAR, N. et A. ANDREOU. 1998, «Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition», *Speech Communication*, vol. 26, nº 4, p. 283–297. 7, 56, 75

LACHHAB, O., J. D. MARTINO, E. H. ELHAJ et A. HAMMOUCH. 2012, «Real time context-independent phone recognition using a simplified statistical training algorithm», *3rd International Conference on Multimedia Computing and Systems - ICMCS'12*. URL https://hal.inria.fr/hal-00761816/document. 6, 31, 36

LACHHAB, O., J. D. MARTINO, E. I. ELHAJ et A. HAMMOUCH. 2014, «Improving the recognition of pathological voice using the discriminant HLDA transformation», *In third IEEE International Colloquium in Information Science and Technology (CIST)*, p. 370–373. 7, 75, 93

LACHHAB, O., J. D. MARTINO, E. I. ELHAJ et A. HAMMOUCH. 2015, «A preliminary study on improving the recognition of esophageal speech using a hybrid system based on statistical voice conversion», *SpringerPlus*, vol. 4, nº 1, doi:10.1186/s40064-015-1428-2,

p. 1–14, ISSN 2193-1801. URL http://dx.doi.org/10.1186/s40064-015-1428-2. 7, 82, 95

LAMEL, L. et J. GAUVAIN. 1993, «High performance speaker-independant phone recognition using cdhmm», *Proc. Eurospeech*, p. 121–124. 17

LAURES, S. J. et K. BUNTON. 2003, «Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions», *Journal of Communication Disorders,*, vol. 36, nº 6, p. 449–464. 66

LAURES, S. J. et G. WEISMER. 1999, «The effects of a flattened fundamental frequency on intelligibility at the sentence level», *Journal of Speech, Language and Hearing Research*, vol. 42, nº 5, p. 1148–1156. 66

LEE, K. et H. HON. 1989, «Speaker-independent phone recognition using hidden markov models», *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, nº 11, p. 1641–1648.
doi:10.1109/29.46546. 17, 33, 41, 49

LEE, K., H. HON et R. REDDY. 1990, «An overview of the sphinx speech recognition system», *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 38, nº 1, p. 35–45.
doi:10.1109/29.45616. 17, 49

LINDE, Y., A. BUZO et R. GRAY. 1980, «An algorithm for vector quantizer design», *IEEE Transactions on Communications*, vol. 28, nº 1, p. 84–95. 37, 99

LIU, H., Q. ZHAO, M. WAN et S. WANG. 2006, «Enhancement of electrolarynx speech based on auditory masking», *Biomedical Engineering, IEEE Transactions*, vol. 53, nº 5, p. 865–874. 5, 81

LJOLJE, A. 1994, «High accuracy phone recognition using context clustering and quasi-triphone models», *Computer Speech and Language*, vol. 8, nº 2, p. 129–151. 48, 49

LOSCOS, A. et J. BONADA. 2006, «Esophageal voice enhancement by modeling radiated pulses in frequency domain», *In Proceedings of 121st Convention of the Audio Engineering Society, San Francisco, CA, USA*, p. 3–6. 5, 80

MACQUEEN, J., L. M. LECAM et J. NEYMAN. 1967, «Some methods of classification and analysis of multivariate observations», *Proc. 5th Berkeley Symposium on Math., Stat.*, p. 281. 99

MANTILLA-CAEIROS, A., M. NAKANO-MIYATAKE et H. PEREZ-MEANA. 2010, «A pattern recognition based esophageal speech enhancement system», *Journal Applied Research & Tech.*, vol. 8, nº 1, p. 56–71. 81

MARKEL, J. D. et A. H. GRAY. 1976, «Linear prediction of speech», *Springer, Communication and Cybernetics*, vol. 12.
doi:10.1007/978-3-642-66286-7. 15

MATUI, K., N. HARA, N. KOBAYASHI et H. HIROSE. 1999, «Enhancement of esophageal speech using formant synthesis», *Proc. ICASSP,*, vol. 1, p. 1831–1834. 5, 80

MELTZNER, G. 2003, *Perceptual and Acoustic Impacts of Aberrant Properties of Electrolaryngeal Speech*, thèse de doctorat, PhD thesis, Massachusetts Institute of Technology. 67

MING, J. et F. J. SMITH. 1998, «Improved phone recognition using bayesian triphone models», *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, p. 409–412. 17

MOULINES, E. et F. CHARPENTIER. 1990, «Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones», *Speech communication journal*, vol. 9, nº 5, p. 453–467. 90

NAKAMURA, K., T. TODA, H. SARUWATARI et K. SHIKANO. 2012, «Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech», *Speech Communication journal*, vol. 54, nº 1, p. 134–146. 83, 90

NARENDRANATH, M., H. MURTHY, S. RAJENDRAN et B. YEGNANARAYANA. 1995, «Transformation of formants for voice conversion using artificial neural networks», *Speech Communication journal*, vol. 16, nº 1, p. 207–2016. 88

NING, B. et Q. YINGYONG. 1997, «Application of speech conversion to alaryngeal speech enhancement», *IEEE Transactions on Speech and Audio Processing*, vol. 5, nº 1, p. 97–105. 6, 82, 83, 84, 90

NORMANDIN, Y., R. CARDIN et DE RENATO MORI. 1994, «High-performance connected digit recognition using maximum mutual information estimation», *Speech and Audio Processing, IEEE Transactions on*, vol. 2, nº 2, p. 299–311. doi:10.1109/89.279279. 20

OHTANI, Y., T. TODA, H. SARUWATARI et K. SHIKANO. 2006, «Maximum likelihood voice conversion based on gmm with straight mixed excitation», *Proc. Interspeech.*, p. 2266–2269. 94

DEL POZO, A. et S. YOUNG. 2006, «Continuous tracheoesophageal speech repair», *Proc. EUSIPCO*, p. 1–5. 80

DEL POZO, A. et S. YOUNG. 2008, «Repairing tracheoesophageal speech duration», *Proc. Speech Prosody*, p. 187–190. 81

PRAVENA, D., S. DHIVYA et A. DURGA DEVI. 2012, «Pathological voice recognition for vocal fold disease», *International Journal of Computer Applications*, vol. 47, nº 13, p. 31–37. 60, 74

QI, Y. et B. WEINBERG. 1991, «Low-frequency energy deficit in electrolaryngeal speech», *Journal of Speech and Hearing Research*, vol. 34, nº 6, p. 1250–1256. 66

RABINER, L. 1989, «A tutorial on hidden markov models and selected applications in speech recognition», *Proceedings of the IEEE*, vol. 77, nº 2, p. 257–286. doi:10.1109/5.18626. 17, 21

REHAN, K. A., V. M. PRASAD, J. KANAGALINGAM, C. M. NUTTING, P. CLARKE, P. RHYS-EVANS, et K. J. HARRINGTON. 2007, «Assessment of the formant frequencies in normal and laryngectomized individuals using linear predictive coding», *Journal of Voice*, vol. 21, nº 6, p. 661–668. 67

ROBINSON, A. 1994, «An application of recurrent nets to phone probability estimation», *Neural Networks, IEEE Transactions on*, vol. 5, nº 2, p. 298–305. doi:10.1109/72.279192. 17

ROBINSON, T. et F. FALLSIDE. 1991, «A recurrent error propagation network speech recognition system», *Computer Speech and Language*, vol. 5, nº 3, p. 259–274. doi:10.1016/0885-2308(91)90010-N. 17

RUMELHART, D. E., G. E. HINTON et R. J. WILLIAMS. 1986, «Parallel distributed processing : Explorations in the microstructure of cognition, vol. 1», chap. Learning Internal Representations by Error Propagation, MIT Press, Cambridge, MA, USA, ISBN 0-262-68053-X, p. 318–362. URL http://dl.acm.org/citation.cfm?id=104279.104293. 88

SAKOE, H. et S. CHIBA. 1971, «A dynamic programming approach to continuous speech recognition», *Proc. 7th Int. Congr. on Acoustics, Budapest, Hungary*, vol. 11, p. 65–68. 29, 84

SHARIFZADEH, H. R., I. V. MCLOUGHLIN et F. AHMADI. 2010, «Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec», *Biomedical Engineering, IEEE Transactions*, vol. 57, nº 10, p. 2448–2458. 81

SIOHAN, O. 1995, «On the robustness of linear discriminant analysis as a preprocessing step for noisy speech recognition», *International Conference on Acoustics, Speech, and Signal Processing ICASSP*, vol. 1, p. 125–128 vol.1. 56

STYLIANOU, Y. 1996, *Harmonic plus noise models for speech, combined with statistical methods for speech and speaker modifications*, thèse de doctorat, ParisTech ENST, Paris, France. 91

STYLIANOU, Y., O. CAPPÉ et E. MOULINES. 1998, «Continuous probabilistic transform for voice conversion», *IEEE Proc. on Speech and Audio Processing*, vol. 6, nº 2, p. 131–142. 83, 84, 86, 88, 89, 91, 95, 99

TANAKA, K., T. TODA, G. NEUBIG, S. SAKTI et S. NAKAMURA. 2014, «A hybrid approach to electrolaryngeal speech enhancement based on noise reduction and statistical excitation generation», *IEICE Transactions on Information and Systems*, vol. E97-D, nº 6, p. 1429–1437. 6, 82, 83, 90, 93, 94

TEBELSKIS, J. 1995, *Speeh Reognition using Neural Networks*, thèse de doctorat, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania. 17

TODA, T. 2003, *High-quality and flexible speech synthesis with segment selection and voice conversion*, thèse de doctorat, School of Information Science, Nara Institute of Science and Technology, Japan. 95

TODA, T., W. BLACK et K. TOKUDA. 2007, «Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, nº 8, p. 2222–2235. 84, 90, 93, 105, 108

TODA, T., K. NAKAMURA, H. SEKIMOTO et K. SHIKANO. 2009, «Voice conversion for various types of body transmitted speech», *In Proc. ICASSP*, p. 285–288. 83

TODA, T., Y. OHTANI et K. SHIKANO. 2006, «Eigenvoice conversion based on gaussian mixture model», *Proc. ICSLP*, p. 2446–2449. 108

TOKUHIRA, M. et Y. ARIKI. 1999, «Effectiveness of kltransformation in spectral delta expansion», *Eurospeech 99*, p. 359–362. 55

TURK, O. et L. ARSLAN. 2006, «Robust processing techniques for voice conversion», *Computer Speech Language journal*, vol. 4, nº 20, p. 441–467. 91

TÜRKMEN, H. et M. KARSLIGIL. 2008, «Reconstruction of dysphonic speech by melp», *Lecture Notes in Computer Science*, vol. 5197, p. 767–774. 80

VALBRET, H., E. MOULINES et J. TUBACH. 1992, «Voice transformation using psola technique», *In Proc. ICASSP*, p. 145–148. 84, 91

VAPNIK, V. 1998, *Statistical Learning Theory*, Wiley, ISBN 978-0-471-03003-4. 17

VINTSYNK, T. K. 1968, «Speech discrimination by dynnmic programming», *Kibernetika (Cybernetics)*, vol. 4, nº 1, p. 81–88.
doi:10.1007/BF01074755. 29

VITERBI, A. 1967, «Error bounds for convolutional codes and an asymptotically optimum decoding algorithm», *Information Theory, IEEE Transactions on*, vol. 13, nº 2, p. 260–269.
doi:10.1109/TIT.1967.1054010. 27

WERGHI, A., J. D. MARTINO et S. B. JEBARA. 2010, «On the use of an iterative estimation of continuous probabilistic transforms for voice conversion», *in Proceedings of the 5th International Symposium on Image/Video Communication over fixed and Mobile Networks (ISIVC)*, p. 1–4. 84, 99

WILPON, J., C. LEE et L. RABINER. 1993, «Connected digit recognition based on improved acoustic resolution», *Computer Speech and Language*, vol. 7, p. 15–26. 17

WUYTS, L., M. S. DE BODT, G. MOLENBERGHS, M. REMACLE, L. HEYLEN, B.MILLET, K. VAN LIERDE, J. RAES et P. H. VAN DE HEYNING. 2000, «The dysphonia severity index : an objective measure of vocal quality based on a multiparameter approach», *In Journal of Speech, Language, and Hearing Research*, vol. 43, nº 3, p. 796–809. 60, 75

YINGYOUNG, Q. 1990, «Replacing tracheoesophageal voicing sources using LPC synthesis», *Journal of the Acoustical Society of America*, vol. 88, nº 1, p. 1228–1235,. 80

YOUNG, S., D. KERSHAW, J. ODELL, D. OLLASON, V. VALTCHEV et P. WOODLAND. 2006, *The HTK Book Revised for HTK Version 3.4*. 7, 31, 40, 75

YOUNG, S., N. RUSSEL et J. THORNTON. 1989, «Token passing : a simple conceptual model for connected speech recognition systems», *Technical Report CUED-Speech Group, Cambridge. web*. 44

YOUNG, S. J., J. J. ODELL et P. C. WOODLAND. 1994, «Tree-based state tying for high accuracy acoustic modeling», *Proc. ARPA Workshop Human Language Technol.*, p. 307–312. 49, 50

YOUNG, S. J. et P. C. WOODLAND. 1994, «State clustering in hmm-based continuous speech recognition», *Computer Speech and Language*, vol. 8, nº 4, p. 369–384. 48, 49

YU, P., M. OUAKINE, J. REVIS et A. GIOVANNI. 2001, «Objective voice analysis for dysphonic patients : a multiparametric protocol including acoustic and aerodynamic measurements», *In Journal Voice,*, vol. 15, nº 4, p. 529–542. 60, 75

ZWEIG, G. et S. RUSSELL. 1999, «Probabilistic modeling with bayesian networks for automatic speech recognition», *Australian Journal of Intelligent Information Processing*, vol. 5, nº 4, p. 253–260. 17